Essays on Regulatory Design

David Thompson

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2021

Abstract

Essays on Regulatory Design

David Thompson

This dissertation consists of three essays on the design of regulatory systems intended to inform market participants about product quality. The central theme is how asymmetric information problems influence the incentives of customers, regulated firms, and certifiers, and the implications these distortions have for welfare and market design.

The first chapter, *Regulation by Information Provision*, studies quality provision in New York City's elevator maintenance market. In this market, service providers maintain machines and are inspected periodically by city inspectors. I find evidence that monitoring frictions create moral hazard for service providers. In the absence of perfect monitoring, buildings rely on signals generated by the regulator to hold service providers accountable, cancelling contracts when bad news arrives and preserving them when good news arrives. Regulatory instruments, such as inspection frequency and fine levels, can therefore influence provider effort in two ways: (i) by directly changing the cost of effort (e.g. fines for poor peformance); (ii) by changing expected future revenue (through building cancellation decisions).

Using a structural search model of the industry, I find that the second channel is the dominant one. In particular, I note that strengthening the information channel has two equilibrium effects: first, it increases provider effort; and second, it shifts share towards higher-quality matches since buildings can more quickly sever unproductive relationships. These findings have important policy implications, as they suggest that efficient information provision — for example, targeting inspections to newly-formed relationships — is a promising avenues for welfare improvement.

The second chapter, *Quality Disclosure Design*, studies a similar regulatory scheme, but emphasizes the incentives of the certifier. In particular, I argue that restaurant inspectors in New York City are locally averse to giving restaurants poor grades: restaurants whose inspections are on the border of an A versus a B grade are disproportionately given an A. The impact of this bias is twofold: first, it degrades the quality of the information provided to the market, as there is substan-

tial heterogeneity in food-poisoning risk even within A restaurants. Second, by making it easier to achieve passing grades, inspector bias reduces incentives for restaurants to invest in their health practices. After developing a model of the inspector-restaurant interaction, counterfactual work suggests that stricter grading along the A-B boundary could generate substantial improvements in food-poisoning rates.

The policy implications of these findings depends on the source of inspector bias. I find some evidence that bias is bureaucratic in nature: when inspectors have inspection decisions overturned in an administrative trial, they are more likely to score leniently along the A-B boundary in their other inspections. However, it's not clear whether this behavior stems from administrative burden (a desire to avoid more trials) or a desire to avoid looking incompetent. Pilot programs that reduce the administrative burden of giving B grades are a promising avenue for future research.

The last chapter, *Real-Time Inference*, also studies the incentives of certifiers, namely MLB umpires charged with classifying pitches as balls or strikes. Unlike in *Quality Disclosure Design*, I find that umpire ball/strike decisions are remarkably bias-free. Previous literature on this topic has noted a tendency for umpires to — for a fixed pitch location — call more strikes in hitter's counts and more balls in pitcher's counts. I propose a simple rational explanation for this behavior: umpires are Bayesian. In hitter's counts, such as 3-0, pitchers tend to throw pitches right down the middle of the plate, whereas in pitcher's counts, they throw pitches outside the strike zone. For a borderline pitch, the umpire's prior will push it towards the strike zone in a 3-0 count and away from the strike-zone in an 0-2 count, producing the exact divergence in ball/strike calls noted in previous work. While implications for broader policy are not immediately obvious, I note several features of the environment that are conducive to umpires effectively approximating optimal inference, particularly the frequent, data-driven feedback that umpires receive on their performance.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I am deeply grateful to my Columbia advisors Andrea Prat and Kate Ho. I would also like to thank all the Columbia faculty and students who provided valuable guidance on early drafts and presentations of my work: Mike Riordan, Matt Backus, Cailin Slattery, Tobias Salz, Steve Olley, Nate Mark, and the rest of the Columbia Industrial Organization Colloquium. I would also like to acknowledge and thank David Love of Williams College for introducing me to economics and encouraging me to pursue the subject in graduate school.

Finally, and most notably, I am ever grateful to my wife Courtney for her unwavering support, encouragement, and patience throughout this process. Thank you for quarantining with me and the cats while this dissertation took shape.

Chapter 1: Regulation by Information Provision

1.1 Introduction

Imperfectly observed product quality is a common source of asymmetric information problems. Unraveling results such as Grossman 1981 and Milgrom 1981 argue that firms will voluntarily disclose their quality, but it is not difficult to generate environments in which such results fail.[1] Indeed, many of us rely on quality disclosure mechanisms other than voluntary firm disclosure every day, whether it be through informal mechanisms (prior experience with firms); third-party reports (e.g. Consumer Reports or S&P ratings); or government-mandated disclosures (e.g. SEC filings or local restaurant inspections).[2]

To what extent disclosure mechanisms improve social welfare is an active area of research. The theoretical and empirical literatures have pointed out several challenges to designing effective regimes. First, disclosers' incentives may not align with those of a social planner.[3] Second, the information supplied to consumers may be redundant or ignored.[4] Third, firms may game the mechanism, for example by directing attention away from unmeasured dimensions of quality.[5] Unsurprisingly, empirical studies of quality disclosure mechanisms report varied results.[6]

[1] For example, if disclosure is costly, only sufficiently high-quality firms may disclose; see Grossman and Hart 1980 and Jovanovic 1982.

[2] See Dranove and Jin 2010 for a comprehensive review of the quality disclosure literature.

[3] Perhaps the most active area in this literature is the incentives of ratings agencies leading up the financial crisis. See, for example, Farhi et al. 2013.

[4] For an example of redundant information, see Dranove and Sfekas 2008, which shows that cardiovascular surgery report cards did not have a strong effect on market share among well-regarded hospitals.

[5] See, e.g., Bar-Isaac et al. 2012.

[6] Many studies find that consumer demand is sensitive to disclosures, (for example, Hastings and Weinstein 2008 on school choice; Jin and Sorensen 2006 on health plan choice; and Cabral and Hortaçsu 2010 on the impact of eBay reviews on demand). Evidence of substantial quality gains from disclosure policies is less abundant. Jin and Leslie 2003 find a roughly 20% reduction in food poisoning cases in response to mandated hygiene cards in Los Angeles, but other studies find evidence of the gaming noted above. Werner and Asch 2005 note that cardiac surgeon report cards may have led surgeons to turn down sick patients, while Jacob and Levitt 2003 find that some teachers and administrators are incentivized to cheat the standardized testing system.

The empirical literature on quality disclosure has largely focused on settings where there is little to no interaction between buyer and seller prior to the transaction — most of us will only choose where to receive surgery or attend school a handful of times in life. In such a setting, disclosure serves as a stand-in for a firm's reputation that would develop naturally in a repeated game. In addition, the existing literature has largely focused, in the words of Dranove and Jin 2010, on "vertical sorting," or settings in which consumers agree on the meaning of higher quality. This chapter aims to fill some of these gaps by asking whether quality disclosure matters when firms form long-term relationships with their customers and the discloser primarily reveals "horizontal" information.

In particular, I study the government-mandated disclosure program in New York City's elevator maintenance market. New York's Department of Buildings ("DOB") is charged with inspecting machines annually and fining buildings and their service providers when unsatisfactory conditions are present. This is an attractive setting to study the questions above for three reasons: first, buildings tend to work with their service providers for many years, so it is not obvious that the DOB is providing any information buildings are not already aware of. Second, service provider quality may be both horizontally and vertically differentiated. Buildings and their elevators are heterogeneous, and certain service providers may have more expertise with particular machines, repairs, or maintenance practices than others. Given the stochastic nature of elevator failures, it may not always be known ex-ante how strong a fit a particular service provider will be. Third, the DOB's match-specific inspection data helps tease apart punitive effects of inspections (fines) from information effects (reputation and contract formation).

To begin, I argue that asymmetric information problems lead buildings and service providers to contract ineffectively over quality. I first note that service provider effort varies depending on the building they are serving. To show this, I construct a proxy for provider effort, their "procedural citation" rate. Procedural citations are violations for simple items — such as providing a maintenance logbook — whose completion is plausibly uncorrelated with the state of the building's machines. Procedural citations are more common in buildings with few elevators, in residen-

tial buildings, and in non-commercially owned buildings. This heterogeneity is consistent with providers exerting less effort in buildings where monitoring is limited, but could also reflect differences in the desire for elevator maintenance across buildings.

I then show the DOB inspections have sizable effects on building behavior. Using missed DOB inspections as an instrument, I estimate that a DOB violation has a causal effect of increasing a building's probability of switching service providers by 4.5 percentage points. The logic behind the instrument is simple: if inspections are missed at random, buildings with missed inspections are a valid control group. So long as inspections only influence switching decisions through the presence of a violation, the causal effect of a violation can be inferred by comparing the difference in switching rates between treatment and control to the difference in violation frequency.

It is difficult to rationalize the influence of DOB violations if buildings and providers contract on a desired level of quality. By revealed preference, the expected utility of working with a provider falls after a DOB violation. This could be due to a decrease in the expected utility the building receives from the work on the machine, or an increase in expected regulatory costs. The fact that the DOB's inspection process is Markovian — there is no formal, nor empirically observed, mechanism for escalating scrutiny on buildings that receive a violation — suggests the former mechanism is driving the contract cancellation response. Moreover, since buildings tend to switch to higher-quality providers following a violation, I argue that the increase in contract cancellation is due to worsened beliefs about provider quality.[7]

In order to more precisely understand the sensitivity of service providers to the regulatory environment, I develop a structural model of the industry.[8] In the first stage, the regulator announces their policy instruments, which include an inspection probability and a fine level. In the second stage, providers set their effort levels and then enter a search process with buildings. The regulator periodically inspects the provider's work and levies fines when violations are detected. Buildings

[7] A subtler interpretation is that DOB violations inform buildings about their machines, and they then search out service providers that are the best match for their particular elevator. This is a multidimensional analog to the quality updating logic above, and generates similar reputational incentives for providers.

[8] For other works that takes a structural approach to regulatory analysis, see Wolak 1994; Timmins 2002; Gagnepain and Ivaldi 2002; Ryan 2012; Lim and Yurukoglu 2018; and Abito n.d.

react to these signals, and decide whether to retain or end their service contract.

The model is an extension of the quality-provision search model of Galenianos and Gavazza 2017, which focuses on the contracting problems between dealers (service providers in this setting) and buyers (buildings) inherent in the illicit drug market. I extend their model in two ways: first, I relax the assumption that quality is perfectly observable after consumption, allowing buildings to learn from regulatory signals in a Bayesian fashion. Second, I incorporate two dimensions of service provider quality: a publicly known component (vertical differentiation), and a match-specific component unknown to the building or provider (horizontal differentiation). This setup captures the intuition that some service providers (e.g. Otis) are known to be high quality, but that there are still match-specific uncertainties.

The model implies an elasticity of aggregate service provider quality with respect to regulatory fines and inspections of 0.34 and 0.15, respectively.[9] All else equal, a 1% increase in the fine level generates the same increase in penalties as a 1% increase in the inspection rate, but additional inspections also changes buildings' information sets. The difference in the elasticities, therefore, represents the informational effects of the regulation. Aggregate quality is more sensitive to inspections than fines for two reason: first, the increased probability of being caught and losing their contract raises the marginal cost of shirking for service providers. Second, since buildings receive information about their service providers faster, they can more quickly determine whether a particular service provider is a good match they should retain, or a bad match they should move on from.

I lastly use the model to estimate outcomes under alternative regulatory schemes. In particular, I show that a scheme in which regulators inspect younger relationships more frequently could reduce aggregate violations by 4.1 percentage points without requiring additional inspections. The intuition behind this result is simple: building beliefs about a provider's unobserved quality are most sensitive to information at the beginning of the relationship. Providing more information early on helps improve aggregate quality by facilitating better matches more quickly.

[9] "Aggregate quality" is the share-weighted violation rate across all service providers. See Section 1.7.

The results of the structural model relate to a growing literature on regulatory discretion. Some authors, such as Stigler 1971, are suspicious of "rich" regulatory regimes, as they may provide leeway for bureaucrats to pursue their private ends at the expense of social welfare. However, recent empirical work has highlighted that there are real benefits to allowing regulators to condition their policies on observable information. One benefit of discretion is the ability to target enforcement towards bad actors: Duflo et al. 2018 argue that emissions inspections in the Gujarat region of India effectively target the worst polluters, in a manner that achieves more abatement than a uniform inspection policy would. In addition, Blundell et al. 2020 find that, in a context with heterogeneous plants and an inability to contract on plant type, dynamic escalation of regulatory scrutiny achieves greater pollution abatement than a uniform policy. Another benefit of regulatory discretion is the ability for regulators to pursue improvements where they are most valued; for example, Kang and Silveira 2020 find that wastewater inspectors in California effectively tailor their enforcement to reflect local preferences for clean water as well as discharger compliance costs.

This chapter notes a distinct benefit of regulatory discretion: the asymmetric nature of information. Even if service providers were homogenous (other than their horizontal match quality), increased scrutiny in the early stages of building/provider relationships would still create the match quality benefits described above.

The remainder of the chapter is organized as follows. Sections 1.2 describes New York City's elevator inspection program. Section 1.3 describes the data sources and provides reduced form evidence for service provider shirking and building learning. Sections 1.4 to 1.6 develop and estimate a structural model of the industry. Section 1.7 presents the counterfactual analysis, and Section 1.8 concludes.

1.2 New York's Elevator Inspection Program

Elevator Service Contracts

In the United States, most elevator buildings employ a third-party service provider to maintain their machines. The goal of elevator maintenance is to keep machines running safely and reliably.

Since it is expensive to replace machines, the average elevator is maintained for 20-25 years before being replaced.[10] The question of how much maintenance a machine should receive, and for how long, is similar to the bus engine replacement problem studied by Rust 1987, with optimal provision depending on the cost of maintenance, the cost of replacement, and the benefit of improved safety and reliability.

Maintenance quality has important welfare implications: McCann 2013 estimates elevators and escalators are responsible for 17,000 injuries and dozens of deaths annually, of which "many could have been prevented if adequate maintenance and inspection procedures had been in place." A more pervasive, but harder to measure, effect of poor maintenance is increased wait times, especially for disabled or elderly passengers that are reliant on elevators for building access.

Because dangerous situations, like entrapments, can arise suddenly, New York City requires buildings to have a service contract in place with an approved provider.[11] Service contracts are similar to insurance contracts: in exchange for a monthly fee, the provider performs periodic maintenance on the machine, identifies necessary repairs, and is on-call in the event of a breakdown. Since many repairs are covered under these contracts, maintenance payments comprise 80-90% of the revenue in this market, making incentives somewhat distinct from those in diagnosis and repair markets considered by previous work such as Darby and Karni 1973 and Schneider 2012.

With some exceptions, maintenance contracts describe the tasks the provider will perform, but do not commit specific mechanic hours or visits.[12] A representative contract template states "the Contractor shall examine the equipment at regular intervals sufficient to preserve the life of the equipment."[13]

Why Regulate at All?

Conversations with industry participants suggest that contracting frictions and moral hazard on the part of service providers can lead to suboptimal maintenance provision. In such an environ-

[10] See Elevator Source.
[11] See DOB Guide, p. 2.
[12] At more complex sites, like airports or skyscrapers, service providers may place mechanic(s) on site permanently.
[13] See McCormick PCS n.d., p. 25.

ment, informal contract enforcement mechanisms need not produce first- or second-best outcomes, leaving room for well-designed regulation to be welfare-enhancing.

Not committing to specific visit dates gives service providers valuable flexibility — if a machine has broken down, the manager can divert a mechanic from his preventative maintenance work to address the breakdown. However, this contractual vagueness can create moral hazard if it is difficult to monitor or enforce claims of insufficient quality. Since labor constitutes nearly 90% of maintenance costs, service providers have strong incentives to reduce mechanic hours however possible.

Several industry commentators point out this monitoring problem, with one maintenance how-to guide advising: "companies that offer really low bids may require more supervision to ensure that they're not skimping on their obligations."[14] Says another commentator in Canada: "A shocking number of elevator contractors, both large and small, are currently not providing the frequency of visits that they have contracted to provide...it is the building owner's responsibility to ensure that the legislated level of service is being provided."[15]

A classic result from principal-agent theory is that the optimal compensation scheme for a risk-neutral agent facing moral hazard is to make them the residual claimant over their output.[16] This scheme is difficult to implement in the elevator industry: the benefits of good maintenance are diffuse and hard to capture or measure. Even if a satisfactory measure were available, the cumulative nature of elevator maintenance means that missed work today may not manifest itself until well into the future, allowing current providers to exert an externality on future service providers by skimping on maintenance today.[17]

In the absence of complete liability for the service provider, buildings enforce their service contracts through a mixture of informal channels (e.g. relational contracts) and formal channels (the courts). Courts are mostly relied on in severe circumstances such as personal injury, but

[14] See Kroll 2017.

[15] See Guderian 2014.

[16] See, for example, Armstrong and Sappington 2007, Section 2.6.1.

[17] This issue is prevalent enough that in most service contracts, the prospective service provider will inspect machines prior to taking control of them and provide the building with a list of repairs that must be addressed before the provider is willing to assume responsibility for the machine.

existing case law highlights how the cumulative nature of maintenance makes it difficult to find service providers liable for insufficient service. When machines break or occupants are injured, courts try to determine whether the failure was due to the actions of the service provider or if the failure would have occurred regardless. For example, in Medinas vs MILT Holdings LLC[18] — a case in which a machine with a DOB "cease use" order fell three stories, injuring a passenger — the court found that even negligent inspections would be insufficient for finding a service provider liable:

> Even accepting for purposes of this analysis that [the defendant] negligently inspected the elevator...and negligently failed to correctly assess the condition of the elevator and necessary repair...it cannot be said to have launched a force or instrument of harm. That is, in failing to correctly inspect or repair the elevator, it did not create or exacerbate an unsafe condition.

While I am not arguing this service provider should have been found liable, this is a useful example to highlight the difficulties in assigning liability when dealing with a dynamic, cumulative process like maintenance.

In the presence of imperfect monitoring and enforcement, buildings may rely on informal channels to induce good behavior on the part of their service providers. The logic is that of relational contracts described in Baker et al. 2002: buildings can take advantage of the repeat nature of maintenance and punish providers who shirk by cancelling the contract in the future. In Section 1.3.2 I provide evidence that buildings do indeed employ this tactic.

Department of Buildings Overview

In New York City, service providers and buildings are regulated by the Department of Buildings (DOB). The goal of the unit is to ensure the "operational safety, reliable service and lawful use of vertical transportation devices throughout our City."[19] The Department pursues this goal through

[18] See *Medinas vs MILT Holdings LLC*, New York State Supreme Court, Appellate Division, July 2015 (131 A.D.3d 121 (N.Y. App. Div. 2015)).

[19] See `https://www1.nyc.gov/site/buildings/safety/elevators.page`, last accessed December 13, 2020.

periodic inspections of the city's elevators. The department's inspectors also follow up on elevator complaints, ensure violations are corrected, and conduct surveys for newly installed elevators.

The DOB requires each machine in the city be inspected twice per year, once during a "routine" inspection which consists of a visual inspection of the machine and the machine room, and once during a "category" inspection where the machine is observed while running. Routines inspections are performed by DOB inspectors or by DOB-licensed third parties, while category inspections are typically done by the building's service provider. Despite the DOB's stated goals, not every machine receives a routine inspection each year. I estimate 81% of machines received a routine inspection between 2008 and 2016.

Inspection criteria are known to market participants, and conform to standards developed by the American Society of Mechanical Engineers.[20] Where violating conditions are found, the DOB has authority to fine the offending service provider depending on the severity of the violation. Fine amounts are highly standardized; a breakdown of the DOB's violation categories and fines is shown in Table 1.1.

Variation in DOB Policy

While most of the variation exploited in this chapter is cross-sectional, I do rely on some time-series variation to identify service providers' cost of quality function. In particular, I use a July 2008 policy change which increased fines and inspection frequency. The increased scrutiny came from a few channels. First, the DOB increased overall inspection frequency. As mentioned above, approximately 81% of machines received an annual routine inspection between 2008-16, an increase from 73% in 1996-2008. Second, the DOB increased fine levels. As shown in Table 1.1, the average fine amount jumped almost 60%, from $382 to $595.[21]

Lastly, the DOB changed its standards for passing category tests. Previously, the building's service provider performed category tests and signed off on them with little review from the De-

[20]The forms used to conduct tests are available online, for example `https://www1.nyc.gov/assets/buildings/pdf/elv3ins.pdf`

[21]A direct comparison of fine levels is difficult to make because severity categories changed in 2008 as well, but fines for the most severe violations, for example, increased 25%. The higher prevalence of more severe violations after 2008 is due to how the new violation categories were defined; the underlying conditions cited are largely unchanged.

partment. Following a rise in violations leading up to 2008, the Department suspected providers were underreporting issues during their category tests, and instituted a requirement that tests be witnessed by an independent third-party.[22] This increased the financial burden on service providers for two reasons: one, pass rates for category tests dropped substantially after the introduction of witnessing (see Table 1.3); and two, the DOB instituted a new penalty that fined buildings for failing to confirm correction following category test failures. These "ACC1" penalties carry a $3,000 fine and quickly became a substantial source of income for the DOB (see Table 1.1).

While witnessing requirements clearly impacted the outcomes of category tests, it is unclear how to map category test outcomes to provider quality. I instead base quality comparisons on DOB violations filed through routine inspections, as those provide a consistent measure of quality over time. I can then estimate aggregate regulatory costs as a function of provider quality, and — as outlined in Section 1.5 — back out the marginal cost of quality by observing how provider quality responds to changes in the regulatory costs.

[22] See Arieff 2009.

Table 1.1: ECB and DOB Penalty Schedule

	1996 - June 2008		July 2008-Present	
	Standard fine	% of violations	Standard fine	% of violations
ECB Violations				
Hazardous	$800	7.1%		
Non-Hazardous	$350	92.9%		
Class 1			$1,000	22.1%
Class 2			$500	72.8%
Class 3			$200	4.8%
Avg standard fine	$382		$595	
DOB Violations				
General elevator	None	86.9	None	74.8
CAT1 Filing	$1,030	13.1	$3,000	12.2
CAT1 Correction			$3,000	12.0
CAT5 Filing			$5,000	0.9
Avg non-general fine	$1,030		$3,072	

Description: Regulatory fee structure from 1996-present. "Standard fines" may not represent actual fines, due to write-offs and markdowns.

1.3 Data and Reduced Form

1.3.1 Data

My data comes primarily from the Department of Buildings and covers several dimensions of the regulatory process: (i) a machine-level record of every elevator inspection performed by the DOB; (ii) a detailed history of every violation levied by the DOB; (iii) a history of customer-generated 3-1-1 complaints. I also collect detailed building-level characteristics.[23]

Inspection and violation records are largely complete extending back to the early 1990s, while complaint data is sparse prior to 2005. Inspections track the date, the type of inspection, a high-level outcome (e.g. pass/fail), and who the performing inspector was. Since 2009, the witnessing agency for category tests has also been recorded. For violations and complaints, I know the location

[23]Building characteristics are collected from the DOB via their Job Application Filings dataset: see NYC Open Data Job Filings. From these applications I extract the building's location, size, ownership type, and its occupancy type. When discrepancies for any field arise across job applications, I use the modal reported value. I supplement this with a DOB machine database posted to Kaggle (see Kaggle 2015) to calculate the number of elevators present in each building.

Table 1.2: Overview of NYC Building Data

	Manhattan	Brooklyn	Queens	Bronx	Staten Island
No. Buildings	14,833	7,629	5,177	4,187	662
Elevators/Building	2.9	1.7	1.8	1.7	1.7
Height	111.8	62.1	59.1	76.1	41.9
No. Stories	10.2	5.6	5.3	6.8	3.5
Ownership					
% Non-Corp owned	48.6	56.3	47.8	52.7	54.8
% Corp owned	47.8	34.7	45.5	36.9	34.6
% Govt owned	3.7	9.1	6.6	10.4	10.7
Occupancy					
% Residential occ.	66.8	66.9	58.8	72.6	39.8
% Commercial ccc.	20.8	12.3	16.8	8.0	30.7
% Other occ.	12.4	20.9	24.4	19.3	29.5
No. Machines					
% 1 machine	49.9	65.9	59.9	61.3	66.9
% 2 machines	22.8	24.1	26.8	27.8	22.2
% 3 machines	9.0	4.4	6.3	4.3	5.1
% 4+ machines	18.3	5.6	7.0	6.6	5.7

Description: Descriptive statistics for the primary sample.

and time of the event, the offending machine, a text description, a measure of severity, and any financial penalty imposed.

I restrict the sample to buildings that contain elevators, and violations and complaints that are elevator-specific. In all, I have information on approximately 70,000 machines spread across 30,000 buildings in the five boroughs of New York. Table 1.2 gives an overview of the sample.

I also use the inspection data to generate two additional fields: first, I calculate for each machine whether it received a routine inspection in a particular year. Second, I construct a yearly mapping between service providers and buildings, which allows me to track provider turnover over time and in response to regulatory violations.[24]

[24]This was done using category tests, which are primarily performed by the building's provider. For each building-year, I assigned the provider performing the majority of tests as that building's provider. This is not always possible (either because the building used a non-service provider to perform the test, or because of ties), but 72% of building-years can be assigned this way. To increase coverage, I filled in some data gaps. For example, if a building employs provider A in year N and year $N+2$ but I am unable to determine the provider in year $N+1$, I assign provider A to

Table 1.3: Overview of NYC DOB Inspections

	1996 - June 2008	July 2008 - 2016
Active Machines	66,936	74,736
% w/Routine Inspection	72.9%	80.7%
% w/CAT1 Inspection	84.5%	83.5%
% Satisfactory	72.6	32.5
% Unsatisfactory	27.4	67.5
ECB Violations		
No. per year	8,197	5,718
% Most Severe	7.1	22.1
Avg. Fine	$474.3	$790.5
Total Fines	$3.89M	$4.52M
3-1-1 Complaints		
No. per year	10,784	9,787
% A priority	3.6	3.4
% Resulting in violation	32.1	48.9

Description: Overview of regulatory outcomes from 1996 to 2016.

I supplement the DOB data with contract financials for a single service provider from 1996-2016. These contracts contain revenue and cost data, which are useful for identifying several of the parameters in the empirical model. See Section 1.5 for more details.

Table 1.3 gives a summary of regulatory inspections since 1996. I focus on routine inspections and Category 1 inspections since they represent about 75% of the inspections in the data. Contrary to the DOB's stated policies, not every machine receives a routine inspection every year. I estimate 73% of machines received inspections during the first regulatory period, although that number has increased to 81% in recent years. The rate of Category 1 completions has remained steady at around 85%, but the initial pass rate on these tests has plummeted since the introduction of mandated witnessing in 2008, as discussed in Section 1.2.

There are two shortcomings of the data. The first is that pricing data is available for only one provider, which requires me to rely on a pricing equation to identify the model. The second is that I do not have easy access to an outcome measure like machine downtime, number of entrapments,

year $N + 1$. I fill such gaps up to two years, which boosts coverage to 78% of building years.

or injuries, which makes welfare analysis difficult. In the main analysis I measure output in terms of DOB violations.

1.3.2 Reduced Form Evidence of Shirking and Learning

Shirking

There is significant variation in building outcomes. Table 1.4 shows that 3-1-1 complaints and ECB violations are far more common in residential buildings, non-corporately owned buildings, and buildings with one or two machines. There could be several explanations for these differences, such as regulator attention, worse quality machines, or worse quality service.[25] As discussed in Appendix A.3, I do not find evidence that the regulator targets inspections by building type, so the primary question is whether Table 1.4 represents differences in machine quality or provider performance.

To disentangle machine quality from provider effort, I restrict attention to what I call "procedural citations." These are components of an ECB violation that should be independent of the machine's underlying state. For example, citations can be given if a machine does not have a proper NYC ID, or if the machine room is missing a maintenance log, or if the machine room was left unlocked. For these violations, any service provider who spent time with the machine should be able to address them in short order. I claim procedural citations are a proxy for service provider effort.

Table 1.5 regresses procedural citations at a machine-year level on building characteristics and provider fixed effects. As before, small, residential, non-corporately owned buildings are more likely to suffer procedural defects. Since we're controlling for provider fixed effects and procedural errors should only reflect differences in provider effort, this suggests that providers exert less effort in these kinds of buildings.

The fact that service provider effort is endogenous need not be too surprising: buildings in New

[25] Complaints, for example, clearly cluster in residential buildings. This need not represent differences in quality, however, as non-residential buildings are much more likely to have maintenance staff on hand to act as a first line of defense. For this reason I prefer using DOB violations as measures of quality.

Table 1.4: Complaint and Violation Incidence

	3-1-1 Complaints	ECB Viols	% Machines
Total/year	9,787	5,717	
By Location			
% Manhattan	29.4	37.0	58.5
% Brooklyn	23.9	23.7	17.5
% Queens	13.9	16.1	12.6
% Bronx	31.4	22.1	9.9
% Staten Island	1.5	1.2	1.5
By Occupancy			
% Residential	91.5	78.4	56.2
% Commercial	4.5	12.4	26.6
% Other	4.0	9.2	17.2
By Ownership			
% Non-Corporate	62.1	57.9	47.2
% Corporate	35.4	39.2	46.2
% Govt	2.4	2.8	6.6
By No. Machines			
% 1 machine	52.1	43.5	26.9
% 2 machines	27.4	27.9	21.7
% 3 machines	7.3	8.0	9.0
% 4+ machines	13.2	20.5	42.4

Description: Complaint and violation incidence since July 1, 2008.

York vary substantially in their use of elevators and their utility from high performance. However, in the next subsection I argue that this divergence in effort stems in part from moral hazard on the part of service providers.

Learning

Before turning to the theoretical model, I note that the regulator appears to provide useful information to buildings through its inspections. In Table 1.6, I regress buildings' turnover decisions on a host of regulatory outcomes and building characteristics. A few things stand out. First, the presence of violations or complaints is strongly correlated with the turnover decision. In particular, the first complaint or violation seems to be most meaningful. For example, the presence of one

Table 1.5: Procedural Citation Incidence

	coef	std err	t	$P > \|t\|$	[0.025	0.975]
Intercept	0.199	0.014	14.489	0.000	0.172	0.226
No. Machines	-0.001	0.000	-9.499	0.000	-0.001	-0.001
Occupancy (Excl. = Commercial)						
Residential	0.0543	0.002	22.250	0.000	0.049	0.059
Other	-0.0314	0.003	-9.481	0.000	-0.038	-0.025
Owner (Excl. = Corporate)						
Non-corporate	0.0243	0.002	12.086	0.000	0.020	0.028
Govt	-0.002	0.006	-0.265	0.791	-0.014	0.011
Borough (Excl. = Bronx)						
Brooklyn	-0.084	0.005	-18.470	0.000	-0.093	-0.075
Manhattan	-0.114	0.004	-29.475	0.000	-0.122	-0.107
Queens	-0.090	0.005	-19.542	0.000	-0.099	-0.081
Staten Island	-0.015	0.010	-1.534	0.125	-0.034	0.004
R^2	0.04					
N	476,382					
Fixed Effects	Service Provider, Year					

Description: Regression of procedural citations on building characters and provider fixed effects. Unit of observation is a machine-year. 53% of provider FE are significant at 5%. Data is from 2008 onwards due to data availability for procedural codes. Procedural violations are defined as violations whose condition is in {dirty, expired tag, unsecured, device not tagged, unlabeled, padlocked, unlocked} or whose relevant component is in {maintenance log, current one year tag, current five year tag, NYC device #}.

ECB violation raises the probability of cancelling a contract by 1.42 percentage points, whereas subsequent ECB violations raise the probability by only 0.22 points.

Of course, these patterns could be consistent with a world in which the regulator is providing no information at all. For example, if the building can perfectly observe the provider's quality level, then violations and turnover could both be driven by the underlying performance of the provider. Thankfully, routine inspections are a convenient instrument to address this problem. As discussed above, routine regulatory inspections appear to happen at random. Inspections should therefore be independent of the underlying quality of the service provider, but predictive of DOB and ECB violations. Section A.3.2 provides supporting evidence for this exclusion restriction.

Table 1.6: Contract Cancellation Regression

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	9.10	0.32	28.40	0.000	8.47	9.73
No. Machines	0.15	0.02	9.30	0.00	0.12	0.18
DOB Violations > 0	1.61	0.13	12.64	0.00	1.36	1.86
ECB Violations > 0	1.20	0.21	5.67	0.00	0.79	1.62
3-1-1 Complaints > 0	2.09	0.24	8.68	0.00	1.62	2.56
No. DOB Violations	0.10	0.03	3.38	0.00	0.04	0.16
No. ECB Violations	0.22	0.08	2.77	0.01	0.07	0.38
$000 ECB Fines	0.49	0.08	6.40	0.00	0.34	0.64
No. 3-1-1 Complaints	0.16	0.05	3.13	0.00	0.06	0.26
R^2	0.013					
N	347,614					
Fixed Effects	Borough, Ownership, Occupancy, and Year					

Description: Regression of building contract cancellation (1 = cancel) on regulatory outcomes. Unit of observation is a building-year. Units are from 0-100, so regression coefficients indicate percentage point differences in turnover rates.

Table 1.7 reports the results of a regression using percent of machines receiving a routine inspection as an instrument for ECB violations.[26] ECB violations continue to have a significant effect; in fact the total effect has increased to over 4 percentage points for a violation.

It is difficult to rationalize this finding if buildings and providers contract on a desired level of quality. By revealed preference, the expected utility of working with a provider falls after a DOB violation. This could be due to a decrease in the expected utility the building receives from the work on the machine, or an increase in expected regulatory costs. Since the DOB's inspection process is Markovian — there is no formal, nor empirically observed, mechanism for escalating scrutiny on buildings that receive a violation — I suspect the former mechanism is driving the contract cancellation response. Moreover, the fact that buildings tend to switch to higher-quality providers following a violation (see Appendix A.1) suggests that the increase in contract cancellation is due to worsened beliefs about provider quality.

[26]The first-stage is quite strong: going from 0 to 100% inspections is associated with an increase of 0.13 ECB violations ($F > 1000$)

Table 1.7: Contract Cancellation IV

	coef	std err	t	P>|t|	[0.025	0.975]
Pre-2008	9.64	0.22	43.08	0.00	9.20	10.07
Post-2008	9.63	0.17	56.40	0.00	9.30	10.00
No. ECB Violations	4.46	0.89	5.02	0.00	2.72	6.21
No. Machines	0.00	0.05	0.13	0.90	-0.10	0.11
Occupancy (Excl. = Commercial)						
Residential	-0.49	0.16	-3.04	0.00	-0.80	-0.17
Other	-0.82	0.18	-4.47	0.00	-1.18	-0.46
Owner (Excl. = Corporate)						
Non-Corporate	0.65	0.12	5.48	0.00	0.42	0.88
Govt	5.43	0.38	14.31	0.00	4.69	6.17
Borough (Excl. = Bronx)						
Brooklyn	-1.03	0.19	-5.32	0.00	-1.42	-0.65
Manhattan	-0.12	0.19	-0.63	0.53	-0.48	0.25
Queens	0.22	0.20	1.11	0.27	-0.17	0.61
Staten Island	-2.10	0.45	-4.67	0.00	-2.99	-1.22

Description: Note: Unit of observation is a building-year. Units are from 0-100, so regression coefficients indicate percentage point differences in turnover rates.

1.4 A Model of the Inspection Industry

1.4.1 Setup

I consider an infinite horizon, discrete time search model. In the baseline version of the model, a unit mass of service providers compete for the business of buildings. If provider s matches with building b, provider s will maintain their machines for the period. At the end of the period, the machine can either be in-violation (v) or not (n). The state of the machine is revealed if the regulator inspects the machine, which happens with probability r. If a violation is present, the provider is fined ϕ. Following the regulator's announcement, the building is given the opportunity to retain their service provider, or to search for another one.

I assume a mass of service providers because the elevator service market is quite fragmented, with 80-90 active service providers in a given year. In addition, as entry and exit is limited in this

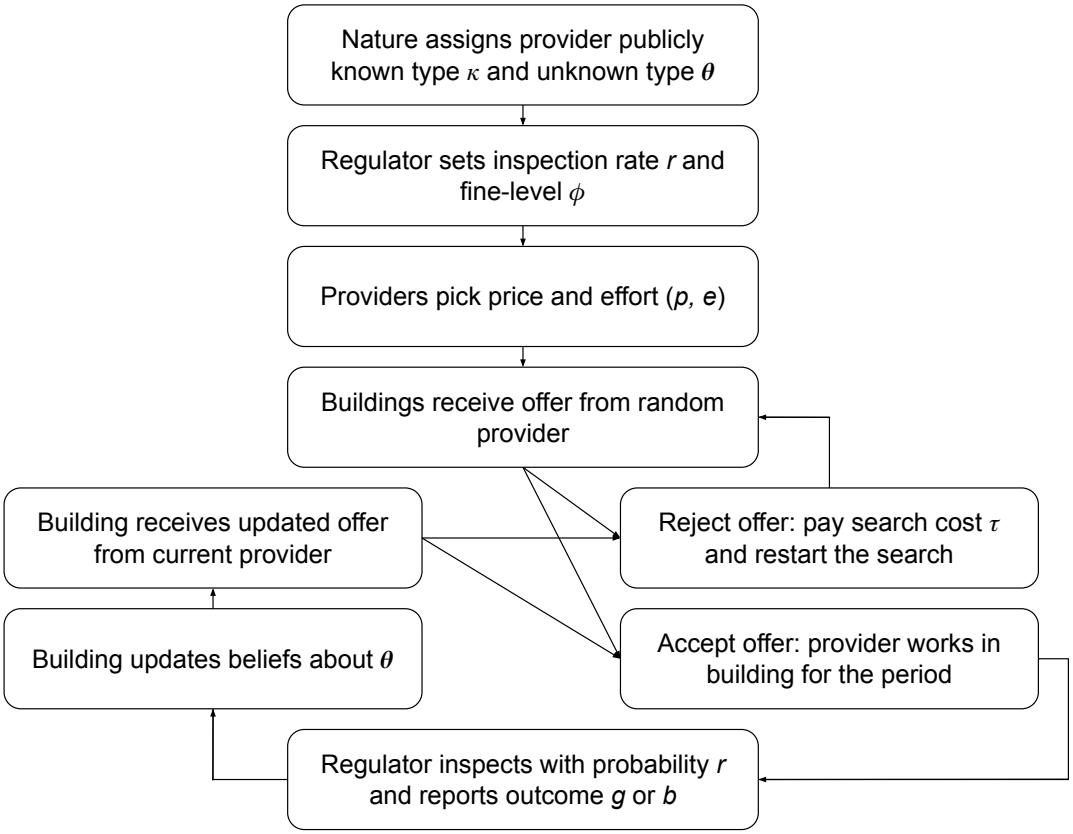


Figure 1.1: Model Overview

market, I do not consider endogenous entry and exit decisions. See Appendix A.3 for more details.

Game Timing

The timing of the game is outlined in Figure 1.1. There are four distinct phases. First, nature assigns providers a publicly known type κ_s and an unknown type θ_{sb}, both of which influence the probability of generating a violation. The regulator then picks its inspection frequency r (the probability it visits a machine in a given period) and fine level ϕ (the amount the service provider is fined when a machine is in violation).

In the third stage, providers pick a price and effort level for each building, (p_{sb}, e_{sb}). These are the prices and efforts provider s will supply if building b chooses to work with them. This setup has two notable features: first, e_{sb} is not observable to buildings, which captures the non-contractibility of quality. However, buildings hold beliefs about effort, $\hat{e}_{sb}$ that satisfy $\hat{e}_{sb} = e_{sb}$ in

equilibrium. This precludes effort from being set at the first-best level e^*_{sb}, since unobservability of quality increases the marginal value of lowering effort. Second, service providers only choose their effort levels once.

In the most general version of this game, providers could alter their effort after each period. I disallow dynamic moral hazard for two reasons. The first is practical: with time-varying effort, the building's learning process (Equation 1.1) must condition on the entire history of observed results, not just the aggregate number of violations. The state space would therefore be exponential in the number of periods, rendering even relatively short time horizons incalculable. Second, dynamic reputation models, such as Holmström 1999, frequently have the feature that incentives to provide effort decline over time as the agent's type becomes more accurately known, which does not appear to be the case in this industry (see Appendix A.2).[27] I instead adopt the view that service provider effort is a fixed outcome of organizational policies and practices, along the lines of Bloom et al. 2017.

In the final stage, buildings engage in random search for providers. In period 0, the building receives a random offer, which consists of a price p_{sb}, the provider's public type, κ_s, and an iid shock ϵ_{sb0}. The building then chooses whether to accept the offer or not. If the offer is not accepted, the building pays a search cost τ and restarts the search process.

The Violation Production Function

Each provider has two types: a publicly known type $\kappa_s \in [0, \bar{\kappa}]$ and an unknown, building-specific type $\theta_{sb} \in [0, 1]$ drawn from distribution F_θ. The public type captures the fact that providers may have well-established reputations for quality, while the unknown type allows for match-specific uncertainty in quality. Buildings will learn about θ_{sb} through the regulator's signals.

Providers also have an endogenous effort e_{sb} which they set at a building level. The probability

[27] Of course, not every dynamic reputation model has this feature. For example, in Mailath and Samuelson 2001, uncertainty about the agent's type is replenished over time, leading to nontrivial dynamics.

that a machine is not in violation depends on effort and the provider's types:

$$P(n|\kappa_s, \theta_{sb}, e_{sb}) = \kappa_s \theta_{sb} e_{sb}$$

The Building's Learning Process

I assume buildings learn about the relationship-specific type θ_{sb} in a Bayesian fashion. In order to do this I need to specify the building's belief about provider s's effort, which I write $\hat{e}_{sb}$. The posterior density of θ after observing n periods of no violations and v periods in violation is:

$$P(\theta|\hat{e}_{sb}, \kappa_s, n, v) \propto (1 - \hat{e}_{sb}\kappa_s\theta)^v \theta^n P(\theta) \tag{1.1}$$

This posterior does not depend on the inspection rate r because I assume buildings observe whether the regulator comes or not. If the inspector does not come, beliefs about θ_{sb} remain unchanged. The inspection rate is important for the building's learning process, though, since it determines the frequency with which the building receives informative signals.

Assumption 1.4.1. *The population distribution of θ can be represented by a density function $P(\theta)$ with support* $[0, 1]$.

Proposition 1.4.1. *For distributions $F(x), G(x)$, let $F \succ G$ denote that F first-order stochastically dominates G. The building's posterior distribution satisfies*

$$P(\theta|\hat{e}, \kappa, n+1, v) \succ P(\theta|\hat{e}, \kappa, n, v) \succ P(\theta|\hat{e}, \kappa, n, v+1)$$

Proof. See Appendix A.4.1. □

This result implies that the expected value of θ is strictly increasing in the number of good signals the building receives.

It is worth noting that the regulator only issues signals about the building's performance in the building. This rules out network effects or social learning in which the building receives signals of

the provider's performance in neighboring buildings.

1.4.2 The Building's Problem

When evaluating an offer from provider s, the building's state variable consists of (dropping b subscripts on variables that are fixed from the building's perspective)

- The provider's price p_s
- The provider's public type κ_s
- The building's beliefs about effort $\hat{e}_s$
- Provider s's regulatory history $\mathcal{I} = \{n, v\}$, where n and v are the number of periods without and with a violation, respectively.
- A random shock ϵ which is observed by the building (but not the econometrician) prior to accepting an offer
- The distribution of offers available in the market $\mathcal{D} = \{p_s, \kappa_s, \hat{e}_s\}_{s\in[0,1]}$.

I do not allow the building to condition on regulatory results of previous providers it has employed; when the building stops working with a provider, I assume it forgets what it has learned about that provider.[28]

Since p_s and κ_s are fixed from the building's perspective, I write the building's expected utility today from matching with provider s as:

$$\begin{aligned} u_s(\mathcal{I}, \epsilon) &= \alpha P_s(n|\mathcal{I}) - p_s + \sigma\epsilon \\ &= \alpha \hat{e}_s \kappa_s \mathbb{E}(\theta|\hat{e}_s, \kappa_s, \mathcal{I}) - p_s + \sigma\epsilon \end{aligned}$$

The parameters α and σ govern the building's taste for quality and the scale of the unobserved shock, respectively. I have scaled the coefficient on prices to 1, meaning utility is measured in dollars.

[28] Given the continuum of providers and random search, this assumption has no effect on the building's payoff. It would if I allowed for directed search, although, since there is no exogenous separation in this model, the providers the building has separated from are all below-average from the building's perspective.

If the building rejects the provider's offer, they pay a search cost τ and return to the search process. The provider's problem is therefore:

$$\begin{aligned} V_s(\mathcal{I}, \epsilon, \mathcal{D}) &= \max\left\{u_s(\mathcal{I}, \epsilon) + \beta \mathbb{E}_s(V_s(\mathcal{I}', \epsilon, \mathcal{D})|\mathcal{I}), \bar{V}(\mathcal{D})\right\} \\ \bar{V}(\mathcal{D}) &= -\tau + \beta \int_s \int_\epsilon V_s(\emptyset, \epsilon, \mathcal{D}) dF(\epsilon) ds \end{aligned}$$

The expectation is taken over the possible regulatory signals that may be generated today and the possible draws of ϵ tomorrow:

$$\mathbb{E}_s(V_s(\mathcal{I}', \epsilon, \mathcal{D})|\mathcal{I}) \equiv \int_\epsilon \left[rP_s(n|\mathcal{I})V_s(\mathcal{I}_n, \epsilon, \mathcal{D}) + rP_s(v|\mathcal{I})V_s(\mathcal{I}_v, \epsilon, \mathcal{D}) + (1-r)V_s(\mathcal{I}, \epsilon, \mathcal{D})\right] dF(\epsilon)$$

where $\mathcal{I}_n$ and $\mathcal{I}_v$ denote the information set following a no-violation (in-violation) outcome.

The provider's policy is a mapping from $(\mathcal{I}, \epsilon, \mathcal{D})$ into $\{0, 1\}$, where 1 represents accepting an offer, and 0 rejecting.

The following lemma is useful for computing equilibria of the dynamic game:

Lemma 1.4.1. *Consider the associated dynamic program defined by:*

$$V_s(\mathcal{I}, \epsilon) = \max\left\{u_s(\mathcal{I}, \epsilon) + \beta \mathbb{E}_s(V_s(\mathcal{I}', \epsilon, \mathcal{D})|\mathcal{I}), \bar{V}\right\}$$

where $\bar{V}$ is an exogenous scalar. If $\bar{V}$ satisfies

$$\bar{V} = -\tau + \beta \int_s \int_\epsilon V_s(\emptyset, \epsilon) dF(\epsilon) ds$$

then V_s solves the original dynamic program as well.

Proof. The proof is immediate, as the equations characterizing the two solutions are equivalent. □

Lemma 1.4.1 makes computing solutions to this dynamic game substantially easier. Providers only care about other providers' actions to the extent they influence $\bar{V}$, and since there are a con-

tinuum of providers, providers take $\bar{V}$ as fixed when they consider changing their prices or effort. Therefore, I can simply fix a $\bar{V}$, find equilibrium strategies for each provider and the building, and then check whether they induce an outside option consistent with $\bar{V}$.

Assumption 1.4.2. *The random shock ϵ can be represented by a full-support density function satisfying $\mathbb{E}(\epsilon) = 0$ and $\mathbb{E}(\epsilon^2) = 1$.*

The mean and variance assumptions are WLOG through rescaling the outside option and σ.

The following proposition summarizes the main results for the building's dynamic program.

Proposition 1.4.2. *For any outside option $\bar{V}$ and effort beliefs $\hat{e}_s$, a unique value function $V_s(\mathcal{I}, \epsilon, \hat{e}_s, \bar{V})$ exists. The building's optimal policy is a cutoff rule such that offers are accepted if and only if $\epsilon \geq \epsilon_s^*(\mathcal{I}, \hat{e}_s, \bar{V})$. The cutoffs are characterized by the indifference condition*

$$\bar{V}(1-\beta) = \alpha P_s(n|\mathcal{I}, \hat{e}_s) - p_s + \sigma\epsilon_s^*(\mathcal{I}) + \sigma\beta\mathbb{E}_s\left[\max\{\epsilon - \epsilon_s^*(\mathcal{I}'), 0\}\right] \tag{1.2}$$

where $\mathbb{E}_s$ is taken over the distribution of ϵ and this period's regulatory outcomes.

Moreover, the value function V_s (and the cutoff ϵ_s^*) are:

- *Weakly decreasing (strictly increasing) in price*
- *Weakly increasing (strictly decreasing) in good signals*
- *Weakly decreasing (strictly increasing) in bad signals*

Proof. See Appendix A.4.2. □

Proposition 1.4.2 establishes that the building is less likely the retain contracts as the number of bad signals increases, and more likely to retain contracts as the number of good signals increases. This is consistent with the reduced form evidence shown in Figure 1.2.

1.4.3 The Provider's Problem

I assume providers maximize the lifetime value of a contract, which is equal to their per-period profits times the expected length of the relationship when first making an offer.

The expected current-period payoff for provider s if they are chosen by the building and have relationship-specific productivity shock θ is:

$$\pi_s(p, e|\theta) = p - c(e) - r\phi(1 - e\kappa_s\theta)$$

Where $c(e)$ is the cost of effort.[29] The final term represents the expected fines provider s will have to pay the regulator.

Let $\ell_s(p, e|\theta, \hat{e}_s, \mathcal{I}, \bar{V})$ denote the expected length of provider s's relationship with the building given the noted state variables. Calculating ℓ_s is complex because the probability of termination varies at every outcome node. Appendix A.4.3 details how to compute ℓ_s and establishes the following comparative statics:

- Contract length is strictly increasing in effort and the number of no-violation periods
- Contract length is strictly decreasing in price and the number of violation periods

Proof. See Appendix A.4.3. □

The provider's problem is therefore to maximize their long-run payoff:

$$p_s^*(\hat{e}_s, \bar{V}), e_s^*(\hat{e}, \bar{V}) = \max_{p,e} \mathbb{E}_\theta \left[\pi_s(p, e|\theta)\ell_s(p, e|\theta, \hat{e}_s, \emptyset, \bar{V})\right] \tag{1.3}$$

Assumption 1.4.3. *The cost function $c(e)$ satisfies*

- $c(0) = c'(0) = 0$
- $c''(e) > 0$
- $\lim_{e\to\bar{e}} = \infty$

Proposition 1.4.3. *For any $\bar{V}$ and any beliefs $\hat{e}_s$, $p_s^*(\hat{e}_s, \bar{V})$ and $e_s^*(\hat{e}, \bar{V})$ exist and are characterized by the first-order conditions of Equation 1.3.*

Proof. See Appendix A.4.4. □

[29] The assumption of a common cost function is driven by the cost structure of the maintenance industry. About 90% of maintenance costs are from unionized labor, so wage rates are very consistent across providers.

The relevant first-order conditions for the provider are:

$$\mathbb{E}_\theta\left[\ell_s(p,e|\theta,\hat{e}_s,\bar{V}) + \pi_s(p,e|\theta)\frac{\partial \ell_s(p,e|\theta,\hat{e}_s,\bar{V})}{\partial p}\right] = 0$$
$$\mathbb{E}_\theta\left[(-c'(e) + r\phi\kappa_s\theta)\,\ell_s(p,e|\theta,\hat{e}_s,\bar{V}) + \pi_s(p,e|\theta)\frac{\partial \ell_s(p,e|\theta,\hat{e}_s,\bar{V})}{\partial e}\right] = 0$$

In a setting with $\theta = 1$ deterministically, the first equation is the standard Lerner index, while dividing the two FOCs gives:

$$c'(e) = r\phi\kappa_s + \frac{\partial_e \ell_s}{|\partial_p \ell_s|}$$

This says that at the optimum, the marginal cost of effort must equal the marginal savings on regulatory fees plus the "marginal rate of contract length substitution" between effort and price.

1.4.4 Equilibrium

I look for pure-strategy, sub-game perfect equilibria that are symmetric in the public type. In particular, an equilibrium consists of provider prices and effort p_s^*, e_s^*, a building policy $\epsilon_s^*(p_s, \mathcal{I})$, effort beliefs $\hat{e}_s$, and an outside option $\bar{V}$ such that

- The building's policy function solves its dynamic program given the outside option $\bar{V}$ (Equation 1.2)
- Providers choose prices and effort optimally given the building's policy (Equation 1.3)
- Providers of the same κ-type choose the same price and effort
- Building beliefs are consistent with provider effort: $e_s^* = \hat{e}_s$ for all s
- The outside option is consistent with the search process

$$\bar{V} = -\tau + \beta \int_s \int_\epsilon V_s(\emptyset, \epsilon) dF(\epsilon_d s$$

Owing to the difficulty of showing that a consistent effort/belief pair exists for each value of the outside option, I cannot prove that such an equilibrium exists.[30] However, equilibria are easily

[30]The difficulty stems from the fact that the optimal effort correspondence is not convex-valued, so Kakutani does not apply. If providers could employ mixed strategies — along with the condition that $\hat{e}_s$ equal the expected value of e_s over the mixing probabilities — then an equilibrium is guaranteed to exist.

found and verified numerically using the following approach:

1. Fix an outside option $\bar{V}$
    - For each provider, find beliefs $\hat{e}_s$ such that $e_s^*(\hat{e}_s, \bar{V}) = \hat{e}_s$. This is a one-dimensional fixed-point problem on [0,1].
    - Calculate the implied outside option

$$\hat{V} = -\tau + \beta \int_s \int_\epsilon V_s(\emptyset, \epsilon) dF(\epsilon) ds$$

2. Iterate over $\bar{V}$ until $\bar{V} = \hat{V}$

At least for the parameter values implied by my empirical setting, I find no issues with missing equilibria. In addition, the procedure above is fast: on a grid with 50 time periods and providers, it takes about a minute to calculate an equilibrium on a standard personal laptop.

1.5 Identification

The goal of identification is to recover the primitives of the model: (i) the population distributions of θ and κ; (ii) the provider's cost function $c(e)$; (iii) the parameters of the building's problem $\{\alpha, \sigma, \tau\}$; and (iv) the discount rate β.

In this section, I show that, with the exception of the discount rate, these quantities are identified by observed regulatory outcomes, the building's provider choices, time-series variation in regulatory parameters, and pricing data for at least one provider. Since there is little in the data to identify the discount rate β, I follow Galenianos and Gavazza 2017 and set $\beta = 0.95$. I show identification in a setting with one building and a long panel for each provider, and so drop all building subscripts in this section.

I make three assumptions to aid identification.[31]

Assumption 1.5.1. *The ϵ shocks are distributed $N(0, 1)$.*

[31] These assumptions are consistent with Assumptions 1.4.2 and 1.4.3.

Assumption 1.5.2. *The unknown type θ has a Beta(k_1, k_2) distribution, where (k_1, k_2) are parameters to be estimated.*

Assumption 1.5.3. *The cost functions has the form:*

$$c(e) = \frac{Ce^2}{\bar{e} - e},$$

where C and $\bar{e}$ are parameters to be estimated.

Assumptions 1.5.2 and 1.5.3 are not necessary as $c(e)$ and the distribution of θ are non-parametrically identified, as discussed below. However, due to the strenuous data requirements needed for such an approach, I have opted to employ the parametric specifications above.

It is harder to relax the normality assumption in Assumption 1.5.1 (the scale and location assumptions are WLOG). I observe cancellation probabilities, which are related to the building's thresholds via the CDF of ϵ. However, for full identification the model needs the precise values of the thresholds, which requires me to take a stand on the distribution of ϵ.

Identifying the building's policy function

The probability that provider s's contract is cancelled at information set $\mathcal{I}$ is:

$$\Pr(\epsilon \leq \epsilon_s^*(\mathcal{I})) = F_\epsilon(\epsilon_s^*(\mathcal{I})) \tag{1.4}$$

The observed rate of contract cancellations is therefore a consistent estimator of the left-hand side of Equation 1.4. Using Assumption 1.5.1, I can invert F_ϵ to recover the underlying values of ϵ_s^*.

Since I do not observe whether contracts are declined prior to a building working with a provider, $\epsilon_s^*(\emptyset)$ is not identified using this method. However, Section 1.5 shows how to use the building's indifference condition to infer $\epsilon_s^*(\emptyset)$.

Identifying $e_s\kappa_s$ and the distribution of θ

For provider s, denote a violation history $v_s^t = \{v_{s1}, \ldots, v_{st}\}$, where $v_{si} = 1$ if a violation is

found, and 0 otherwise.[32] The probability of observing a violation history v_s^t is

$$\begin{aligned} P(v_s^t) &= \int_\theta P(\theta)P(v_s^t|\theta)d\theta \\ &= \int_\theta \prod_{i=1}^{t} P(a|v_s^i)(1-e_s\kappa_s\theta)^{v_{si}}(e_s\kappa_s\theta)^{1-v_{si}}P(\theta)d\theta \\ &\propto \prod_{i=1}^{t}\int_\theta (1-e_s\kappa_s\theta)^{\sum v_{si}}(e_s\kappa_s\theta)^{t-\sum v_{si}}P(\theta)d\theta) \end{aligned}$$

where $P(a|v_s^i)$ is the probability the contract is accepted after history v_s^i. With a parametric assumption $\theta \sim \theta(k)$ I can form the maximum likelihood estimator in k and $e_s\kappa_s$. With sufficient data I could relax Assumption 1.5.2 and use nonparametric likelihood techniques as in Geman and Hwang 1982.

It is easiest to understand the identification of these parameters in a setting with $r = 1$. Suppose we observe a provider for two periods. Then:

$$\begin{aligned} \text{Pr(no violations in first period)} &= e_s\kappa_s\mu_\theta \\ \text{Pr(no violations in both periods)} &= e_s^2\kappa_s^2(\mu_\theta^2+\sigma_\theta^2) \end{aligned}$$

That is $e_s\kappa_s$ controls a provider's violation frequency across buildings, whereas σ_θ^2 determines the likelihood for violations to cluster within a specific building.

Identifying α/σ and $\epsilon_s^*(\emptyset)$

Differencing the building's indifference condition (Equation 1.2) for provider s and two information sets $\mathcal{I}, \mathcal{I}'$ gives

$$0 = \alpha\Delta P_s(\mathcal{I},\mathcal{I}') + \sigma\Delta\epsilon_s^*(\mathcal{I},\mathcal{I}') + \sigma\beta\Delta C_s(\mathcal{I},\mathcal{I}') \tag{1.5}$$

[32] I ignore periods in which no inspection occurred, as they are not informative about the provider's type.

where C_s is a continuation term:

$$C_s(\mathcal{I}) = \sum_v P_s(v|\mathcal{I}) \int_{\epsilon_s^*(\mathcal{I}_v)}^{\infty} (\epsilon - \epsilon_s^*(\mathcal{I}_v)) dF(\epsilon)$$

Other than α and σ, every term in Equation 1.5 is identified at this point. The parameters α and σ are not separately identified using this approach, but their ratio can be estimated using least-squares techniques.[33]

The ratio α/σ represents the relative importance of quality versus unobservable elements of demand. This quantity is identified by the sensitivity of buildings to news: if small changes in the probability of a good outcome lead to large changes in acceptance probabilities, then α is relatively large compared to σ.

Once α/σ has been identified, Equation 1.5 implicitly defines $\epsilon_s^*(\emptyset)$ with $\mathcal{I} = \emptyset$. After dividing through by σ, the only unknown in this equation is $\epsilon_s^*(\emptyset)$, since $C_s(\emptyset)$ only depends on observable thresholds.[34] Knowing $\epsilon_s^*(\emptyset)$ is important, since the probability an initial contract offer will be accepted influences the provider's contract length function ℓ_s.

Identifying the Cost Function

The provider's first-order condition with respect to effort is:

$$\mathbb{E}_\theta \left[(-c'(e) + r\phi\kappa_s\theta)\, \ell_s(p, e|\theta, \hat{e}_s, \bar{V}) + \pi_s(p, e|\theta) \frac{\partial \ell_s(p, e|\theta, \hat{e}_s, \bar{V})}{\partial e} \right] = 0 \tag{1.6}$$

In Appendix A.5.1, I show that, for a provider with known prices, Equation 1.6 defines a first-order differential equation $c'(e) = f(c, e)$ with the initial condition $c(0) = 0$. With sufficient variation in the equilibrium effort level, the cost function could be estimated non-parametrically

[33] For each provider, I retain each information set that contains at least 750 observations and construct differences between each pair. I then choose α/σ to minimize:

$$\sum_{s,\{\mathcal{I},\mathcal{I}'\}} \left(\frac{\alpha}{\sigma} \Delta P_s(\mathcal{I}, \mathcal{I}') + \Delta\epsilon_s^*(\mathcal{I}, \mathcal{I}') + \beta \Delta C_s(\mathcal{I}, \mathcal{I}') \right)^2$$

[34] Since $\epsilon_s^*(\emptyset)$ will lie in between $\epsilon_s^*(\{g\})$ and $\epsilon_s^*(\{b\})$, once I know the relative importance of quality I can infer exactly where it will land between these two observed thresholds.

using numerical techniques such as Euler's Method.

Since I only observe two time periods, I instead use the parametric form mentioned in Assumption 1.5.3, and invert the first-order condition to solve for C and $\bar{e}$.

In Appendix A.5.1 I analyze the provider's first-order condition in a simpler setting to generate intuition for the relevant patterns in the data. In this simplified setting, I find that the marginal cost of effort is identified by the relative change in price versus effort in response to an increase in r. High-cost providers will react in a relatively price-intensive manner, whereas low-cost providers will adjust more on the effort margin.

Identifying σ

In Appendix A.5.3 I show that $\sigma \frac{\partial \epsilon_s^*(\mathcal{I})}{\partial p}$ is identified from the building's indifference condition and the observed cancellation probabilities. Since the provider's contract length ℓ_s depend on price through its effect on ϵ_s^*, I can therefore identify $\sigma \frac{\partial \ell_s}{\partial p}$.

The provider's first-order with respect to price can be written:

$$\sigma = -\frac{\mathbb{E}_\theta \left[\pi_s(p, e|\theta) \frac{\sigma \partial \ell_s(p,e|\theta,\hat{e}_s,\bar{V})}{\partial p} \right]}{\mathbb{E}_\theta \left[\ell_s(p, e|\theta, \hat{e}_s, \bar{V}) \right]}$$

For any provider for which I have price data, every element of the right-hand side is identified, meaning σ is as well. In particular, I note the implied σ will be larger when profits are large (competition is softened when the majority of the building's utility comes from the random shock), and when expected contract length is short (buildings will prioritize searching for good draws of ϵ when its variance is higher).

Identifying τ and the Outside Option

Once σ and α are identified, the building's outside option $\bar{V}$ is determined by its indifference condition, Equation 1.2. In addition, the search cost τ is identified through the relationship:

$$\bar{V} = -\tau + \int_s \int_\epsilon V_s(\emptyset, \epsilon) dF(\epsilon) ds$$

Subtracting $\beta\bar{V}$ from both sides gives

$$\begin{aligned}\bar{V}(1-\beta) &= -\tau + \int_s \int_\epsilon (V_s(\emptyset,\epsilon) - V_s(\emptyset,\epsilon_s^*(\emptyset)) dF(\epsilon) ds \\ &= -\tau + \sigma \int_s \int_{\epsilon_s^*(\emptyset)}^{\infty} (\epsilon - \epsilon_s^*(\emptyset)) dF(\epsilon) ds\end{aligned}$$

Identifying the distribution of κ_s

Lastly, I need to identify the distribution of κ. The provider's FOC with respect to price is

$$\mathbb{E}_\theta\left[\ell_s\right] = -\mathbb{E}_\theta\left[\pi_s \frac{\partial \ell_s}{\partial p}\right]$$

At this point, everything in this equation is identified with the exception of $p_s - c(e_s)$, which can therefore be inferred from this equation.

The provider's FOC with respect to effort can be written:

$$c'(e_s) = \underbrace{\frac{r\phi\kappa_s \mathbb{E}_\theta\left[\theta\ell_s(p,e,\theta)\right]}{\mathbb{E}_\theta\left[\ell_s(p,e,\theta)\right]}}_{\text{avoided fines}} + \underbrace{\frac{\mathbb{E}_\theta\left[\pi_s(p,e,\theta)\partial_e\ell_s(p,e,\theta))\right]}{\mathbb{E}_\theta(\ell_s(p,e,\theta))}}_{\text{additional profits}}$$

Multiplying through by e_s I see that the right-hand side of this equation is identified, meaning $e_s c'(e_s)$ is identified as well.[35] By Assumption 1.4.3, $ec'(e)$ is a strictly increasing function. Since $c(e)$ has already been identified, e_s is therefore identified as well.

1.6 Estimation and Results

Thresholds

Recall the building thresholds can be extracted from the relationship

$$Pr(\text{contract cancelled}|\mathcal{I}) = F_\epsilon(\epsilon^*(p,\mathcal{I}))$$

[35] See Appendix A.5.1 for the argument of why $e_s\partial_e\ell_s$ is identified.

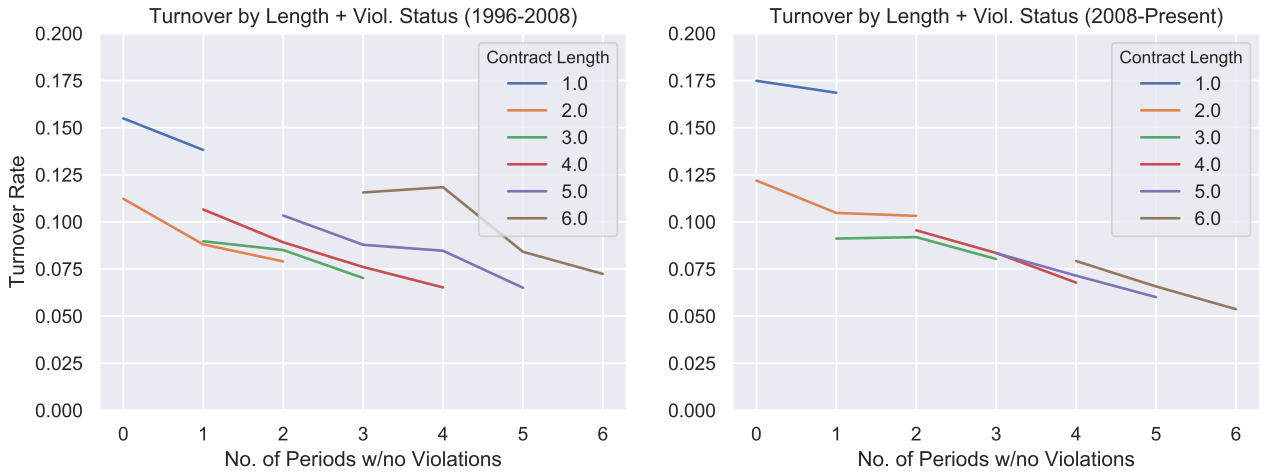


Figure 1.2: Contract Turnover by Violation State

In Figure 1.2 I plot these contract cancellation rates. Each line represents a different contract length, while the x-axis indicates the number of periods of without a violation the contract has experience. For example, the red line in the left chart indicates that four year old contracts that have only had one period without a violation cancel at an 11% rate, whereas those that have never had a violation cancel at a 6% rate.

The model predicts that contracts will cancel more (less) frequently as providers experience bad (good) outcomes. That is, the model makes two predictions about how this chart should look: 1) for a given contract length, the lines should slope downward, as fewer violations lead to lower turnover rates; 2) for a fixed number of no-violation periods, younger contracts should turnover less often than older contracts, since an older contract with the same number of no-violation periods necessarily has experienced more violations.

These predictions are borne out almost perfectly in the pre-2008 data, with the exception that one-year contracts appear to turnover more frequently than expected. Post-2008, one-year contracts continue to cancel at higher than expected rates, and possible two-year contracts as well.

Distribution of κ

Figure 1.3 shows the estimated density of κ using the procedure outlined in Section 1.5.[36] The

[36]In particular, Figure 1.3 shows a histogram of pooled values of κ based on both pre-2008 and post-2008 data.

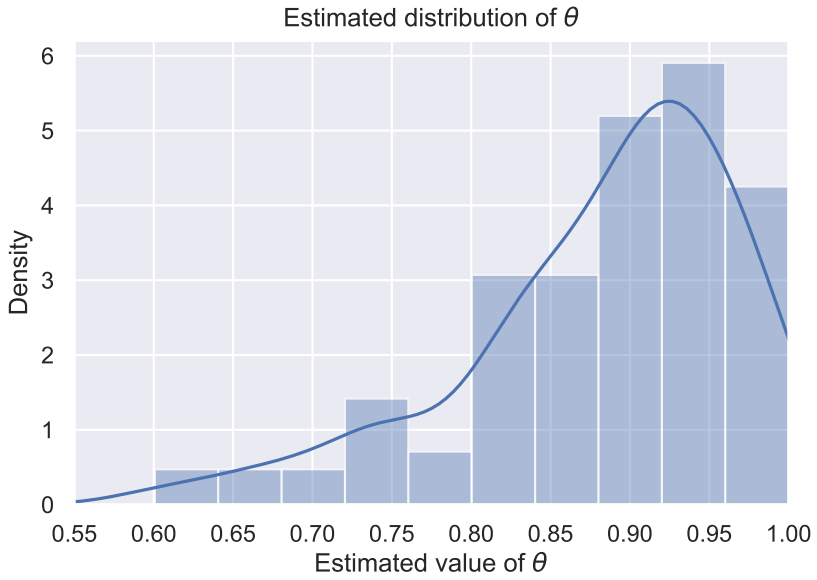


Figure 1.3: Estimated Provider Type κ (pre-2008 data).

distribution appears to be unimodal with a peak between 0.9 and 0.95. The empirical standard deviation of the κ distribution is about 0.1.

My model predicts that κ is fixed, and therefore should not change in response to a shift in regulatory policy. Since I can estimate κ using both pre- and post-2008 data, this allows me to test whether the distribution of κ changes over that time-period. Indeed, since the correlation between a provider's pre-2008 value of κ and its post-2008 value is 0.95, it would appear the estimation is capturing a fixed characteristic of providers. More formally, Figure 1.4 shows a quantile-quantile plot of the pre- and post-2008 distributions. While the post-2008 quantiles are consistently 2-3 percentage points above the pre-2008 ones, a Kolmogorov Smirnov tests fails to reject the null that the two distributions are different from one another at the 5% level.[37]

Parameter Estimates

Estimation follows the sequential identification argument given in Section 1.5. However, in order to increase power, I pool buildings together. To avoid the heterogeneity issues noted in Section 1.3, I report results for all residential, non-corporately owned buildings in New York City

[37]The exact p-value is 0.096.

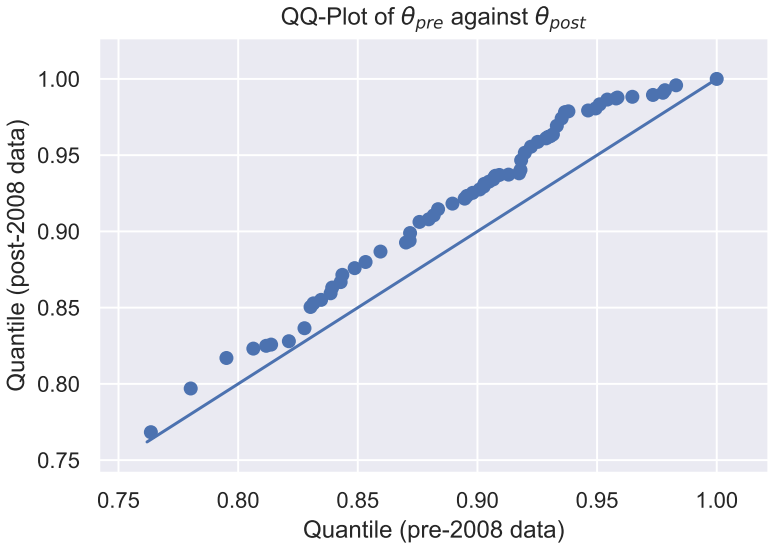


Figure 1.4: Comparison of κ Estimates, pre-2008 vs. post-2008.

with four or fewer machines. Estimates for other groups follow similar patterns.

I generate standard errors via block bootstrap, resampling at the building level. I resample 100 times to generate parameter standard errors. The results of the estimation procedure are reported in Table 1.8.

Coefficient	Estimated Value	Standard Error
α (quality demand)	2.954	0.564
σ (SD of ϵ)	0.325	0.051
τ (search cost)	0.222	0.059
C (cost param)	1.109	0.313
$\bar{e}$ (cost param)	1.401	0.081
θ_1 ($\theta \sim \text{Beta}(\theta_1, \theta_2)$)	9.067	0.552
θ_2	1.00	0.00
Calculated Quantities at Estimated Values		
$\mathbb{E}(\theta)$	0.901	
$\mathbb{E}(\kappa)$	0.962	
$\mathbb{E}\left(\frac{\pi_s}{p_s}\right)$	0.142	

Table 1.8: Estimated Model Coefficients

Parameters α, σ, and τ are measured in thousands of dollars.[38] Thus, a guarantee of no elevator

[38]Given the nonlinearity of the cost function, the units for C and $\bar{e}$ are less meaningful.

violations for a year is worth approximately $3,000 to a building. The unobservable component of service appears to be small, with a one standard-deviation shock being worth $325. I estimate renegotiation costs to be about $200.

1.7 Counterfactuals

This section examines the service market's response to changes in regulatory policy. I first consider changing the inspection rate and the fee level. I then look at the effect of discretionary policies such as targeted inspections. I find that targeted inspections — in particular those that inspect young relationships more frequently — have the potential to reduce aggregate violation prevalence by 4 percentage points without any increase in total inspections. The primary reason for this is that targeted inspections allow providers to quickly identify (and therefore terminate) low-quality matches.

Changing r or ϕ

Before considering more complicated counterfactuals such as regulatory discretion, it is useful to study the effects of a change in the regulator's policies r and ϕ on market outcomes. Let $m(s,\theta)$ be the equilibrium market share (density) of provider s when their unknown type is θ. The provider's aggregate market share is:

$$m(s) = \int_{\theta} m(s,\theta) d\theta$$

Define aggregate market quality as the probability a machine is not in a violation state:

$$\begin{aligned} q &\equiv \int_{s,\theta} e_s \kappa_s \theta m(s,\theta) ds d\theta \\ &= \int_{s,\theta} \theta e_s \kappa_s m(s) \frac{m(s,\theta)}{m(s)} ds d\theta \end{aligned}$$

The elasticity of q with respect to the inspection rate r is:

$$\hat{q}_r = \frac{1}{q} \int_{s,\theta} \theta e_s \kappa_s m(s,\theta) \left[\widehat{e_s} + \widehat{m(s)} + \widehat{\frac{m(s,\theta)}{m(s)}} \right] ds d\theta$$

Figure 1.5: Decomposition of Provider Response to Inspections

Description: Decomposition of the response of aggregate market quality to a 1% increase in r. Both figures show the same data, with the left grouped by "market effects" and the right by "regulatory effects". Bar totals add to 1 and represent the fraction of the change in aggregate quality attributable to each factor.

This highlights that changes in aggregate quality come from three channels:

- Changes in provider effort $\widehat{e_s}$
- Changes in provider market share $\widehat{m(s)}$
- Changes in match quality: $\frac{\widehat{m(s,\theta)}}{m(s)}$

Since, all else equal, the direct cost effect is identical when r or ϕ is raised one percent, any additional changes produced by raising r instead of ϕ are due to the new information environment. This motivates the following decomposition:

$$\hat{q}_r = \underbrace{\hat{q}_\phi}_{\text{Cost effect}} + \underbrace{\hat{q}_r - \hat{q}_\phi}_{\text{Info effect}}$$

Figure 1.5 shows the magnitude of these various effects when r is increased by 1% (from the baseline inspection rate pre-2008). I estimate aggregate quality increases by 0.34%. As shown in the left side of Figure 1.5, roughly half of this increase is due to improvements in "match quality," or the average matched value of θ: as inspection rates increase, buildings recognize poor matches

more quickly and move on from them (and vice-versa for beneficial matches). Interestingly, simply increasing fines also generates some improvement in match quality. This stems from two facts: one, provider effort increases in response to an increase in fines; and two, regulatory signals are more informative about θ when provider effort is high.[39]

Effort effects account for the remainder of the improvement in aggregate quality. Service providers primarily increase their effort due to cost effects: more inspections raise the marginal cost of shirking. However, information effects play an important role here. More inspections increase the number of signals the building receives, and thus the probability of losing a contract if a violation is present.[40]

Lastly, market share effects appear to be relatively small — regulation does not steer the market towards higher-quality providers. This isn't surprising, as the information provided by the regulator has no relationship to the provider's public type κ.

In all, the estimates suggest that information effects account for 60% of the increase in aggregate quality in response to an increase in the inspection rate (see the right side of Figure 1.5).

Regulatory Discretion: Inspecting New Relationships

Given that information is the majority effect of the DOB's inspections, I consider whether there are benefits to alternate methods of information provision. One simple way is to allow inspectors to vary their inspection intensity by the age of a relationship. This kind of discretionary policy may be attractive because it provides more information in the early stages of relationships, allowing bad matches to end quickly and good matches to persist.

I construct such an inspection rule as follows. Define the inspection rate

$$R(t|r,\gamma) \equiv \max\left\{\min\left\{b(r,\gamma) + \frac{\gamma}{1+t}, 1\right\}, 0\right\}$$

[39] Consider the extreme case where $e = 0$. Then every regulatory signal is a violation, and the building never learns about θ.

[40] Again, to take an extreme example, consider the case where $r = 0$. In this case, the building never learns about their service provider, so there is no incentive for the service provider to exert any effort.

where

- $R(t|r,\gamma)$ is the probability a relationship of age t is inspected
- $b(r,\gamma)$ is a baseline inspection rate such that the aggregate inspection probability is r
- γ captures the sensitivity of the inspection rate with respect to relationship length: positive γ means younger relationships are inspected more frequently, and vice versa

Since the regulator's policy conditions on a subset of the building's state variable, this policy does not substantially complicate solving the model of Section 1.4.[41] I simply update the expectation terms in Equations 1.2 and 1.3 to account for the varying inspection probabilities over time.

Given γ and a target inspection rate r, I can calculate the aggregate quality function

$$AQ(r,\gamma) = \int_{s,\theta} \theta e_s(\gamma)\kappa_s m(s,\theta|\gamma) ds d\theta$$

where $e_s(\gamma)$ and $m(s,\theta|\gamma)$ are the equilibrium effort and market shares as a function of γ, respectively.

Finally, for a given inspection rate, I can define the optimal γ and maximum achievable quality as:

$$\begin{aligned} \gamma^*(r) &= \arg\max_{\gamma} AQ(r,\gamma) \\ AQ^*(r) &= AQ(r,\gamma^*(r)) \end{aligned}$$

Figure 1.6 plots $AQ^*(r)$ against the DOB's current policy, $AQ(r,0)$. At the pre-2008 inspection rate, I estimate that setting $\gamma \approx 0.41$ would result in an aggregate quality of 0.901, a 4 percentage point increase over the baseline aggregate quality of 0.86. Under this scheme, new contracts are inspected with probability 1, with the inspection rate gradually falling to 0.59 as $t \to \infty$.

The analysis above also lets me put a simple lower bound on the benefit of discretionary policies. This is useful for evaluating the policy, as a provides a dollar amount against which costs

[41] It is important that the building's posterior on θ does not depend on r. If it did, then the building would have to condition on the entire history of violations, as opposed to the aggregate number of violations.

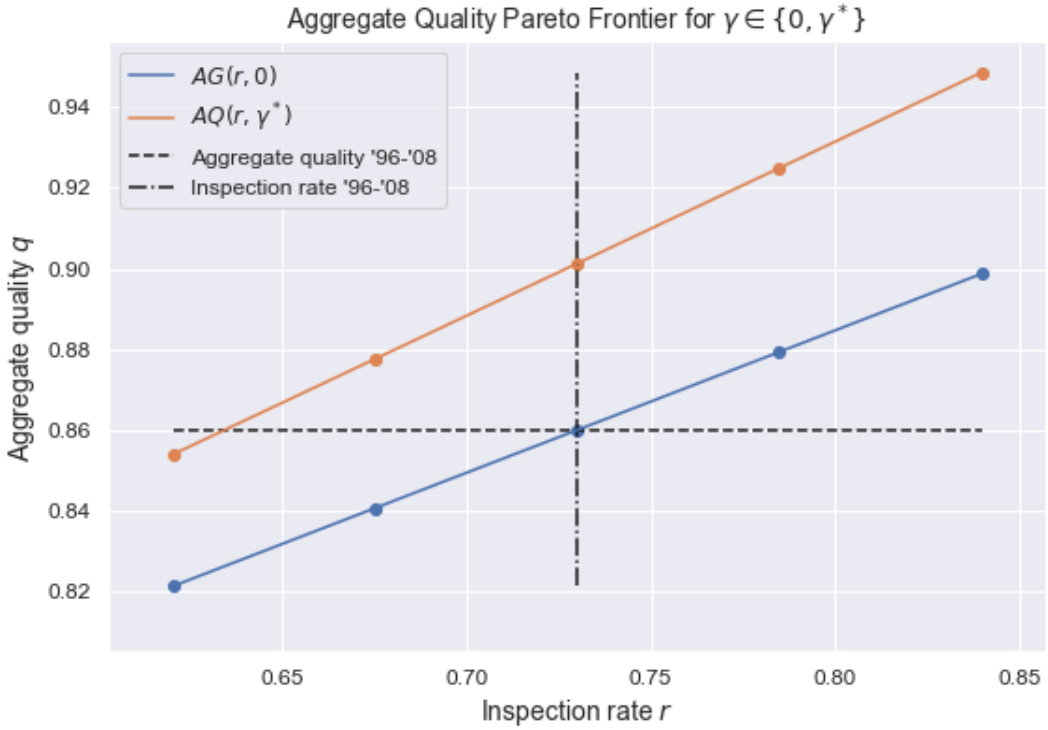


Figure 1.6: Aggregate Quality with Contract Length Targeting

Description: Aggregate quality under (i) optimal contract length targeting ($AQ(r, \gamma^*(r))$, orange line) and no targeting ($AQ(r, 0)$, blue line).

of a more complex policy can be compared. In particular, note that $AQ^*(0.63) \approx AQ(0.73, 0)$. That is, the estimates suggest that, under an optimal discretionary policy, the DOB could reduce its inspection rate by 10 percentage points while maintaining aggregate quality at pre-2008 levels.

1.8 Conclusion

This chapter studies New York City's elevator inspection program, an example of government-mandated quality disclosure. It is not *a priori* clear that such an inspection program is beneficial: buildings and service providers typically work together for many years, and it is unclear if inspectors generate any new information. However, I find reduced-form evidence that inspection results have a sizable impact on contract renewal decisions. Using quasi-experimental variation generated

by missed inspections, I find that a DOB violation has a causal impact of increasing the probability of a contract cancellation by 4.5 percentage points. I find no evidence that this is driven by changes in regulatory behavior following an inspection, suggesting that the information supplied by the regulator is of first-order importance.

I then construct a structural search-model of the elevator industry to simulate behavior under alternate regulatory schemes. I find the elasticity of aggregate industry quality (percent of machines with violating conditions) with respect to the inspection rate is 0.34. Approximately 60% of this effect is due to two information effects: increased inspections allow buildings to more quickly ascertain poor-quality matches, and buildings' increased ability to rely on reputational threats increases the marginal cost of shirking for providers.

Lastly, I find that allowing inspectors to target building-provider pairs based on their relationship length is a promising method for improving aggregate quality. The marginal benefit of information is largest early in a relationship, so front-loading signals allows buildings to more quickly ascertain match quality. I find that within a particular class of targeting schemes, optimal targeting would increase aggregate quality by 4.1 percentage points without requiring additional inspections. Seen another way, under an optimal targeting scheme, the DOB could maintain aggregate quality at pre-2008 levels while decreasing the inspection rate by 10 percentage points.

Chapter 2: Quality Disclosure Design

An Empirical Study of New York's Restaurant Inspection Program

2.1 Introduction

Mandatory disclosure programs, such as financial disclosures, hospital report cards, or local restaurant or building inspections, are frequently used to combat asymmetric information about product quality. In the ideal, disclosure programs produce informed consumption decisions and steer firms towards providing desirable price/quality combinations.

Recent theoretical and empirical work — in addition to the first chapter of this thesis — has shown that proper implementation is crucial to the success of disclosure programs. Failures can occur along the entire chain of consumers, firms, and disclosers. Disclosure may generate information that is irrelevant (as in Dranove and Sfekas 2008) or that goes unnoticed by consumers (as in Dai and Luca 2020 and Makofske 2020b). In addition, firms and disclosers may engage in various types of gaming that degrades the accuracy of the information disclosed.[1] Understanding what features lead to successful disclosure is a matter of active interest.

This chapter asks whether disclosers in government-mandated disclosure programs provide accurate information. While there was substantial interest in the incentives of third-party disclosers in the wake of the 2007-08 financial crisis (see, e.g. Bolton et al. 2012), outside of Makofske 2020a, I am not aware of prior work on the incentives of government-sponsored disclosers. Distortions in this market may be particularly damaging if consumers are more willing to take government-issued information at face value.

I study discloser incentives and the design of New York City's restaurant inspection program.

[1] In multidimensional settings, regulated firms may respond by shifting attention towards measures dimensions of quality and away from unmeasured dimensions (see Lu 2012 and Bar-Isaac et al. 2012). Firms may also engage in outright fraud, such as Jacob and Levitt 2003

I show that New York City restaurant inspectors are disposed towards giving restaurants A grades. This stems in part from the coarse nature of restaurant grading: inspections are given a numerical score (typically between 0-30 points) which is mapped into a publicly posted A, B, or C grade. Payoffs for restaurants and disclosers change discontinuously across grades, raising the concern that manipulation may occur near the boundary. I provide two arguments for the presence of inspector bias, one cross-sectional and one structural.

Every inspection identifies a set of violations, which are assigned points and totaled to give an aggregate score (lower scores are better). Following Makofske 2020a, the cross-sectional evidence for inspector bias relies on differences in inspector behavior during inspections whose observed violations place them near the A-B boundary, or what I call marginal inspections.[2] Since inspectors have discretion in the number of points they assign to a given violation, inspectors disposed towards giving A's should score violations more leniently in marginal inspections. This is exactly what happens: scores bunch on the A side of the A-B boundary even after conditioning on observed violations.

The strongest competing explanation for this data is one of restaurant manipulation: marginal restaurants quickly clean up their act when an inspector comes, reducing the severity of some violations and producing a mass of A grades. There are three reasons why I do not expect restaurant manipulation is sufficient to explain the bunching in the data.[3] First, the reduction in violation scores in marginal inspections is largely due to inspectors assigning zero points to certain violations. This reflects an active choice by inspectors, as recommended point ranges are strictly positive. Second, many common violations are fixed relative to the duration of an inspection, such as construction materials and plumbing. Third, there is no obvious relationship between which violations receive fewer points and which violations would seem to be easily manipulable. For example, inspectors penalize missing food thermometers less harshly in marginal inspections, but it is unclear how restaurants could manipulate the severity of such a binary violation. Alternatively,

[2]See Section 2.4.2 for a definition. This approach is similar to that used by Makofske 2020a in his work on inspector bias among Los Angeles restaurant inspectors.

[3]Since I am conditioning on observed violations, this says nothing about a restaurant's ability to manipulate which violations are observed. I expect restaurants do a lot of this.

personal cleanliness and food temperature violations, which seem easier to manipulate, are scored more harshly in marginal inspections.

Structural Model of Inspector Behavior: The second half of the chapter takes up the question of how inspector behavior impacts welfare. To this end I combine a model of inspector scoring thresholds and restaurant health investments with restaurant-level food poisoning data. This allows me to estimate how inspectors make scoring decisions and how aggregate food-poisoning cases would change given different scoring policies.

The model's identification relies on the panel structure of the data. In particular, I rely on the fact that, regardless of what a restaurant does to prepare for an inspection, there is residual variation in which violations actually occur (e.g. whether a bug walks across the counter, or an employee wipes his hand on the wrong towel while the inspector is looking).[4] Therefore, in the absence of inspector bias, a restaurant's inspection score distribution should vary continuously across the A-B boundary.[5] The model interprets excess A grades near the A-B boundary for a given restaurant as evidence of inspector bias.

Perhaps the biggest contribution of the structural analysis is the inclusion of restaurant-level food-poisoning data. While previous studies, such as Jin and Leslie 2003 and Ho 2012, have used aggregate food-poisoning cases to evaluate the efficacy of grade-card schemes, to the best of my knowledge no studies have used restaurant-level data. While there are concerns about data accuracy and missing data problems (see Section 2.3), this data allows me to construct a detailed picture of how inspector scoring decisions relate to a measure of obvious welfare relevance.

Positively for the Health Department, I find that inspectors are sensitive to food-poisoning risk: riskier restaurants receive worse scores. However, the scores just to the left of the A-B boundary mask a lot of variation in food-poisoning risk. The dirtiest restaurant receiving a score of 13 (the worst A score) is 30% riskier than the cleanest restaurant receiving a 12, a roughly three-fold

[4]The noisiness of restaurant inspections is well-known. Ho 2012 and Makofske 2020a both note that there is little consistency in violation composition from inspection to inspection.

[5]Since inspection scores are quantized there is no literal discontinuity; instead, I model inspector thresholds using an ordered probit with linear heteroskedasticity.

increase in relative risk compared to other adjacent scores. This leniency allows restaurants to reduce their investment in health practices while still receiving A grades.

The primary counterfactual considers the impact of reducing inspector leniency at the A-B boundary. There are two main effects: some restaurants increase their investment in health practices in order to continue receiving A grades, whereas others become discouraged because the investment required to receive A's is too costly. Under current practices, the former effect dominates: I predict that stricter grading at the A-B boundary would result in a 13-22% decline in aggregate food-poisoning cases.

Two further counterfactuals consider the effect of altering institutional details of New York's inspection program, namely re-inspections and preferential inspection timing:

Re-inspections are a second inspection given to restaurants that do not receive an A on their first inspection. Restaurants do not need to post the results of their initial inspection until the re-inspection has occurred. I find that re-inspections increase the frequency of false negatives (dirty restaurants labeled clean) and, for most restaurants, reduce the incentives to invest in health practices due to the relative ease of securing an A grade. I estimate that eliminating re-inspections would reduce aggregate food poisoning cases by 10-14% as well as cut the false-negative rate roughly in half. However, false positive rates would double from about 5% to 10% of inspections, which would likely generate significant opposition from restaurants.

Preferential inspection timing refers to the longer inspection cycles given to restaurants that receive good grades. A restaurant that receives an A on its first inspection will not be inspected again for 11-13 months, whereas a restaurant that receives a C will be re-inspected in 3-5 months. While I agree there are benefits to keeping tabs on dirty restaurants, differentiated inspection timing may actually disincentivize investment in health practices. The intuition for this result is easiest to see by considering an absurd example: suppose A grades lasted for 100 years and B/C grades for one day. So long as their is some noise in the inspection process, restaurants will have minimal incentive to invest in their health practices, as they can be assured that they will soon enjoy the

long-lasting benefits of an A grade.[6] I find that setting equating the length of inspection cycles for A, B, and C grades would reduce food poisoning cases by 10-13%.

It is interesting to contrast the result on preferential timing to those of Blundell et al. 2020, who argue that dynamic escalation mechanisms (e.g. increased scrutiny for repeat offenders under the Clean Air Act) can be an effective regulatory tool when inspections are costly and regulators cannot contract on a firm's type. Preferential timing represents a form of increased scrutiny for poor performers. However, when the regulatory "stick" is a public signal (the letter grade) as opposed to costs directly imposed by the regulator, this form of escalation may actually lead to worse health outcomes due to the possibility of false negatives.

Sources of Inspector Bias: Lastly, I consider possible explanations for regulator bias. I find that scores are more likely to cluster on the A side of the A-B boundary when: (i) the restaurant has never previously been shut down by the health department; (ii) the restaurant is in its first year of operation; and (iii) when the inspector has had recent B grades overturned by the city's Office of Administrative Trials and Hearings (OATH). Moreover, I find that the extent of clustering varies substantially from inspector to inspector.

These findings are consistent with several possible sources of inspector bias. For example, findings (i) and (ii) are consistent with inspectors finding it more costly to give B grades to younger or historically high-performing restaurants. Finding (iii) is consistent with a "minimal squawk" story such as Leaver 2009, where some inspectors may attempt to avoid the reputational harm of mistakes by providing overly generous scores. However, finding (iii) could also be explained by other forces, such as inspector capacity constraints: when inspectors have many cases being overturned by OATH, they are less willing to assign a B grade and risk having to attend another hearing. Clarifying the source of inspector bias is an important next step, as it will help inform policy ideas to mitigate the bias.

The remainder of the chapter is organized as follows. Section 2.2 discusses details of New

[6] In the presence of noisy inspections, restaurants would also have minimal incentives to invest if A grades lasted for one day and B/C grades for 100 years, since sooner or later they would receive a B grade.

York City's restaurant inspection program. Section 2.3 gives details on the inspection and food-poisoning data. Section 2.4 provides reduced form evidence for inspector bias and its sources. Sections 2.5 through 2.7 develop the structural model and estimate various counterfactuals. Section 2.8 concludes.

2.2 Institutional Context

In New York City, the Department of Health and Mental Hygiene ("DOHMH" or "the Health Department") is in charge of inspecting the city's more than 27,000 restaurants. The Health Department employs roughly 100 health inspectors who visit restaurants unannounced and inspect their operations.[7] One useful feature of the inspection program is that inspectors are randomly assigned to restaurants.[8]

The DOHMH groups inspections into *cycles*, which consist of the cycle's *initial inspection* and a potential *re-inspection*.[9] In both of these inspections, health inspectors observe the restaurant's condition and score any violations of the city's health code present at the time of inspection. As shown in Table 2.1, scores vary within and across violation types, with more severe or widespread violations receiving more points. Scores are totaled at the end of the inspection, with lower totals indicating better compliance with the health code.

Restaurants receive a grade card based on their inspection results which they must post in a location "easily seen by people passing by."[10] The relationship between inspection scores and grades is:

- A score of 0-13 corresponds to an "A"
- A score of 14-27 corresponds to an "B"
- A score of 28 or more corresponds to an "C"

[7] In 2016, for example, the Health Department employed 83 inspectors that performed more than 100 inspections, accounting for 96% of all inspections. These inspectors averaged 358 inspections, or one to two per day.

[8] See Krishna 2018.

[9] Restaurants that are not yet open are given a *pre-permit* inspection which they must pass in order for a permit to be issued. I do not include pre-permit inspections in my analysis for two reasons: (i) pre-permit inspections are not graded, and (ii) restaurants only need to pass their pre-permit inspection once, whereas cycle inspections persist in perpetuity.

[10] See Health Department 2011 for details on posting requirements.

Table 2.1: Overview of Common Violations

Violation	Description	Points by Condition Level I	II	III	IV	V
10F	Non-food contact surface improperly constructed. Unacceptable material used, or equipment improperly maintained and/or not properly sealed, raised, spaced or movable to allow for cleaning.	2	3	4	5	—
8A	Facility not vermin proof. Harborage or conditions conducive to attracting vermin to the premises and/or allowing vermin to exist.	—	—	4	5	—
4L	Evidence of mice or live mice present in facility's food and/or non-food areas.	5	6	7	8	28
6D	Food contact surface not washed, rinsed or sanitized after each use and following any activity when contamination may have occurred.	5	6	7	8	—
2G	Cold food item held above 41°F (smoked fish and Reduced Oxygen Packaged food above 38°F), except during necessary preparation.	7	8	9	10	28

Description: Overview of the five most common violations in 2019. The "Condition Level" is a measure of how widespread or severe a violation is. For example, observing one cold food items above 41F would constitute a Condition I for violation 2G, whereas observing four such items would be a level IV. Condition V is reserved for Public Health Hazards, conditions that are immediately dangerous to the public and cannot be remedied while the inspector is onsite.

If the score of an initial inspection exceeds 13, the restaurant need not post a B or C card immediately. Instead, the restaurant is subject to an unannounced re-inspection within approximately one month; if it receives an A on re-inspection, then it is given an A grade card. If it receives a B or C on re-inspection, the restaurant must then post the re-inspection grade.

The Health Department places additional scrutiny on poor-performing restaurants: restaurants that receive an A on the initial inspection will not be inspected again for 11-13 months, whereas those that receive a B or C at any point in the cycle will have their cycle start again in 5-7 and 3-5 months, respectively. Figure 2.1 summarizes the links between inspection scores, restaurant grades, and the timing of the inspection cycle.[11]

Restaurants have the right to challenge inspection results through the city's Office of Admin-

[11] When violating conditions that the Health Department deems a "Public Health Hazard" cannot be corrected during the inspector's visit, the Health Department has the power to close the restaurant immediately. Closures are relatively uncommon, and not my primary focus; there are 2,456 closures in the main dataset, meaning approximately 2% of inspections lead to a closure.

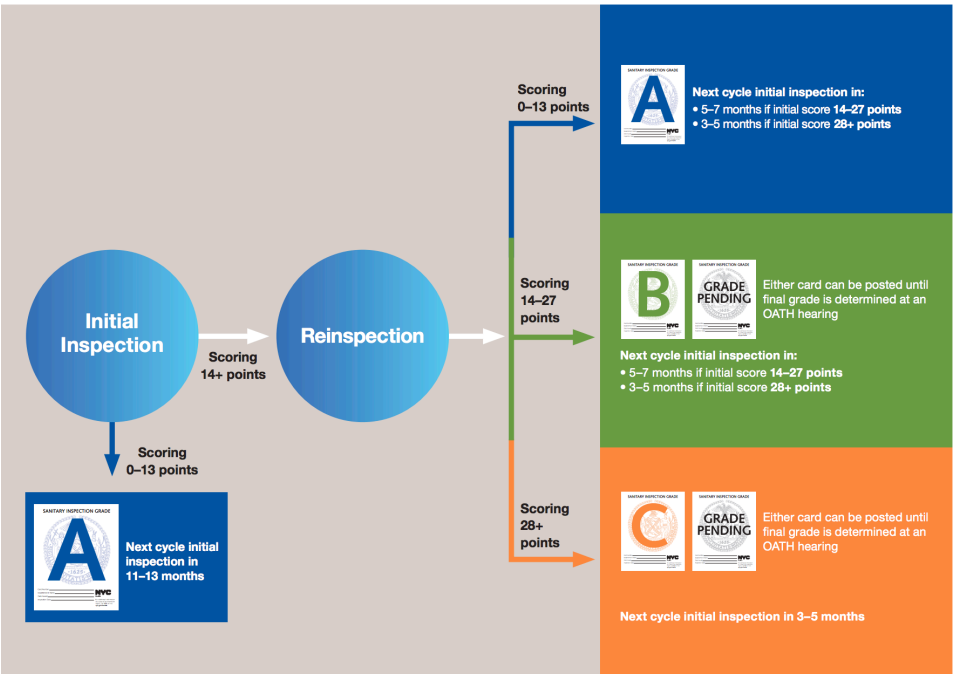


Figure 2.1: New York Restaurant Inspection Cycle

istrative Trials and Hearings (OATH) board. During this time the restaurant is given a "Grade Pending" card, and, following the OATH hearing, the restaurant is given an updated grade card that reflects any violations dismissed by the OATH board. For example, a restaurant that scores 17 on re-inspection would typically receive a B grade. If the OATH board dismisses a 5 point violation, the restaurant's inspection score would be updated to 12, and the restaurant would be given an A grade card.[12] Appendix B.5 shows that OATH procedures are not the driving force behind the score bunching in the data.

In addition to lengthening the inspection cycle, restaurants have a monetary incentive to achieve an A on their initial inspection. Sanitary violations have monetary penalties associated with them, but penalties are waived for restaurants that receive an A on their initial inspection.[13] Restaurants that receive a B or C on their initial inspection are liable for the penalties they incur, even if they

[12]While OATH dismissals can change the grade a restaurant is allowed to post, they do not affect the length of a restaurant's inspection cycle. See Health Department 2010, p. 20.

[13]See Rules of the City of New York, Chapter 23 Appendix C for the penalty schedule and Health Department 2014 for the penalty waiver.

receive an A on re-inspection. As shown in Table 2.2, fines from health inspections generate tens of millions of dollars in revenue for the city annually, costing the average restaurant approximately $1,000.

More importantly, restaurants have a reputational incentive to receive good grades on their inspections. Survey and empirical evidence suggests consumers take grade cards into account when making restaurant choices. Jin and Leslie 2003 estimate that after Los Angeles introduced grade cards in 1998, receiving an A grade led to a 5 percent increase in revenue over a B grade. The average restaurant in New York State generates about $1 million in revenue, so if grades have a similar effect on demand in New York, the average restaurant would stand to lose about $50,000 annually from receiving a B.[14] This is substantially larger than the fines levied by the DOHMH in Table 2.2.[15] In addition to the posted grade cards, the Health Department's inspection scores are available via their website.[16] Moreover, since July of 2018, Yelp has included information generated from restaurant's Health Department inspections on the business's Yelp page.[17] Dai and Luca 2020 and Makofske 2020b find that such information can have strong effects on demand, particularly if the information is presented saliently.

2.3 Summary of Restaurant Inspection Data

The primary data consists of two linked datasets: (i) the last three years of the DOHMH inspection records for all currently operating restaurants in NYC; (ii) all complaints of food poisoning logged through the city's 311 program.[18]

[14] See National Restaurant Association 2019.

[15] In a preliminary paper, Rothbart et al. 2014 find an A grade is associated with a $145 increase in mean daily sales as compared to a B, ceteris paribus. This corresponds to A restaurants earning $52,925 more revenue per year than B restaurants, all else equal.

[16] See `https://a816-health.nyc.gov/ABCEatsRestaurants/#/Search`.

[17] See Lowe 2018.

[18] In addition to the primary inspection and food poisoning data, I supplement the analysis with an older dataset of inspection records from 2007-2016. In addition to the fields noted in the primary inspection dataset, the historical data also includes an inspector identification number as well as information on whether violations were dismissed through an OATH tribunal. In Appendices B.4 and B.5 I use this data to study inspector heterogeneity and the impact of OATH hearings on score clustering.

Despite the additional fields present in the historical data, I use the most recent data as my primary sample. After adopting the restaurant grading scheme in mid-2010, inspection results exhibit strong dynamic trends until stabilizing

Table 2.2: 2019 NYC Health Inspection Violations and Fines.

Violation Category	**No. Viols**	**% of Viols**	**Net Fines**	**Net Fines/Rest.**
General Violations	*50,601*	*41.3*	*$9,164,814*	*$429*
10F: Non-food contact surface improperly constructed	17,417	14.2	$2,237,800	$105
08A: Facility not vermin proof	14,237	11.6	$2,482,200	$116
10B: Plumbing not properly installed or maintained	7,146	5.8	$1,036,200	$49
Critical Violations	*47,887*	*39.1*	*$7,448,200*	*$349*
04L: Evidence of mice	9,560	7.8	$1,698,600	$80
06D: Food contact surface not properly washed	8,377	6.8	$1,185,800	$56
06C: Food not protected from potential contamination	7,056	5.8	$1,195,714	$56
Public Health Hazards	*20,409*	*16.7*	*$4,531,050*	*$212*
02G: Cold food item held above 41F	7,239	5.9	$1,572,500	$74
02B: Hot food item held below 140F	6,068	5.0	$1,301,750	$61
04H: Improper handling of raw, cooked, or prepared food	2,743	2.2	$595,000	$28
Unscored	*3,604*	*2.9*	*$1,714,225*	*$80*
20F: Current letter grade card is not posted	970	0.8	$970,000	$45
22G: Possessing/selling Styrofoam containers	702	0.6	$175,500	$8
20D: "Choking first aid" poster not posted	419	0.3	$73,325	$3
Total	**122,501**	**100**	**$22,858,289**	**$1,070**

Description: Violation groups and top three violations by group for 2019 inspections in the main dataset (which excludes locations that have multiple restaurants at the same address). "Net Fines" is calculated by multiplying each violation by its lowest possible fine level, unless the violation occurred during an initial inspection resulting in an A grade, in which case no fine is assumed. "Net Fines/Rest." is calculated by dividing "Net Fines" by 21,364 (the number of restaurants in the dataset).

2.3.1 Health Department Inspection Records

The Health Department maintains a public database of inspection records for each currently operating restaurant in the city.[19] For each restaurant, results are available for the three years prior to the most recent inspection. Inspection records include:

- The date of the inspection
- The type of inspection (e.g. initial versus re-inspection)

around 2016. Since the goal in this work is not to study these dynamics, but rather the steady-state consequences of such a grading scheme, I opted to focus on the most recent timeframe.

[19] See NYC Open Data Restaurant Inspections.

Table 2.3: Number of Restaurants in Sample by Borough.

	Manhattan	**Brooklyn**	**Queens**	**Bronx**	**Staten Island**	**Total**
Raw Data	10,904	6,852	6,150	2,401	1,002	27,309
Final Data	7,642	5,963	4,945	2,081	727	21,364
% Chains	13.7	10.9	12.3	19.6	13.1	13.2
% American	26.0	18.3	15.4	14.5	23.2	20.3
% Chinese	5.5	10.9	11.6	12.8	8.5	9.2
% Cafe	8.8	6.9	4.3	2.1	2.6	6.4
% Pizza	4.3	4.3	3.6	3.8	4.3	4.1
% Mexican	3.3	4.5	3.5	5.6	5.1	4.0

Description: "Final Data" is the dataset obtained after removing all locations with multiple restaurants at the same BBL and Address combination. Borough totals do not sum to the stated total because 6 restaurants do not have a borough tag.

- Each violation cited
- Any action taken by the Health Department (e.g. a shutdown)
- The inspection score and grade
- Restaurant name, cuisine, and location

I make two alterations to this dataset. First, I add a chain field by tagging restaurants that have 5 or more locations operating under the same name as chain restaurants. Second, I exclude locations that have multiple restaurants operating at the same address (e.g. the Penn Station Plaza in New York) due to difficulties in matching food poisoning reports to the underlying restaurants. Thus the sample represents all stand-alone restaurants in New York City and excludes multi-restaurant locations like food courts. Table 2.3 summarizes the sample.

2.3.2 311 Food Poisoning Reports

Accurate reports of food poisoning are notoriously difficult to collect. In 2011, the CDC estimated that 48 million Americans (one in six) contract foodborne illnesses annually, but only 0.3% of those are hospitalized, meaning the vast majority of food poisoning cases are not reflected in official statistics.[20] I generate restaurant-level measures of food poisoning risk using reports of

[20]See Scallan, Hoekstra, et al. 2011 and Scallan, Griffin, et al. 2011.

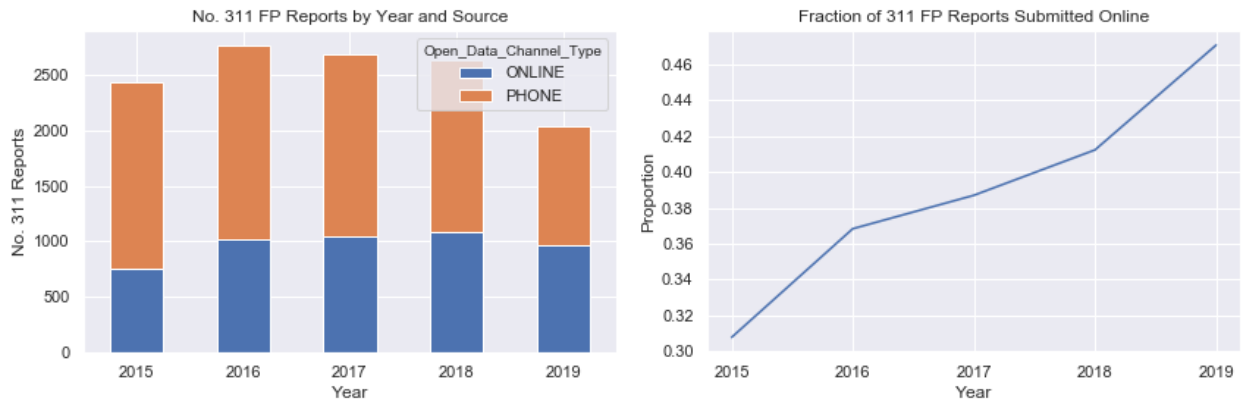


Figure 2.2: Annual 311 Food-Poisoning Reports.

food poisoning to New York's 311 system. Citizens who suspect they have contracted food poisoning from a restaurant can call or report online, and complaints are inputted into a centralized database.[21] I filter for complaints that have occurred since 2015 and that occurred in a "Restaurant/Bar/Deli/Bakery" or "Restaurant". Other location types (e.g. school cafeterias or catering services) are not represented in the inspection database.

Having restaurant-level estimates of risk is a real advance; previous studies on the links between public health policy and food poisoning risk, like Jin and Leslie 2003 or Ho 2012, rely aggregate measurements of food-poisoning. This finer data allows me to better understand how inspectors react to and report the risk present during an inspection.

That said, the 311 data has some notable shortcomings. As shown in Figure 2.2, the 311 system logs on the order of 2500 food poisoning complaints annually. The CDC's one in six ratio implies about 1.5 million cases of food poisoning in New York annually, suggesting the 311 data represents on the order of 1/600th of the food poisoning cases in New York. In addition, DOHMH representatives have stressed that any particular report may be misleading for several reasons: diners may misattribute discomfort from overeating or another illness to food poisoning, or may blame a restaurant for food poisoning when in fact they received it from another source.

The validity of my inferences is therefore dependent on 311 reports of food poisoning being indicative of overall food poisoning risk. In particular, if food poisoning complaints are correlated

[21] See NYC Open Data 311 Requests.

Table 2.4: Summary of 311 FP Report Matching.

	No. FP Reports	Comments
Raw Data	12,633	All 311 reports from 2015-onward whose location type is 'Restaurant/Bar/Deli/Bakery', or 'Restaurant'
Excl. reports w/no address and no BBL	11,474	Caller gave no location information
Excl. reports w/BBL not in the restaurant database	9,059	Mostly grocery stores/delis. Could be restaurants that are now closed
Excl. reports w/ multiple restaurants at address/BBL	6,853	Removing complaints from locations with multiple restaurants
Matched Sample	6,853	

Description: Details of the matching procedure linking 311 food poisoning reports to restaurants.

with the presence of food poisoning, and consumer criteria for lodging 311 complaints remains stable over time, fluctuations in 311 complaints will be indicative of changes in the underlying risk of food poisoning. As I show in Section 2.4, the correlation between 311 incidence and inspection performance suggests the food poisoning data captures meaningful differences in health risks across restaurants.

I use a two-step procedure to match 311 food poisoning reports to the inspection database (since the 311 report does not list the underlying restaurant name). First, I match reports to restaurants using BBL numbers and their address.[22] This fails to identify some matches because many addresses in New York City have multiple valid names (for example, Eight Avenue is also known as Frederick Douglass Boulevard for a stretch near Columbia University). I therefore use a second match criteria that links 311 reports to restaurants if (i) their street number is identical; and (ii) their latitude/longitude coordinates are within one meter of each other. Table 2.4 gives details on the matching process.

[22]BBL = Borough-Block-Lot number, essentially a parcel or building identifier. Large buildings may have multiple addresses associated with a single BBL.

2.4 Reduced Form Evidence

This section investigates the basic patterns in the inspection and food poisoning data. I highlight three main points: (i) inspector grades exhibit noticeable bunching to the left of the A-B boundary; (ii) inspectors appear to be locally averse to giving restaurants non-A grades; (iii) restaurants do not appear to "clean up their act" when an inspection is shortly forthcoming. I lastly look at the determinants of inspector's B-grade aversion, and find that inspectors are more likely to give marginal A scores when they've recently had inspections overturned by OATH hearings.

2.4.1 Score Bunching

Figure 2.3 shows the distribution of inspection scores for initial inspections (left) and re-inspections (right). For initial inspections, there is substantial bunching just to the left of the A-B boundary, which is the point at which a re-inspection is required and restaurants are not given penalty relief. There does not appear to be bunching on the B-C boundary, perhaps due to the fact that little is different for an initial inspection that results in a B versus one that results in a C. For re-inspections, there is bunching both at the A-B boundary and the B-C boundary. All of these patterns are nearly identical over time.[23]

Appendix B.5 investigates whether score bunching is due to restaurants challenging inspection results and having their scores lowered after the initial inspection. The results suggest that while many restaurants successfully challenge B grade inspections, the drop-off along the A-B boundary comes primarily from the initial inspection scores. The presence of score bunching near important regulatory thresholds is not unique; Jin and Leslie 2003 and Makofske 2020a find evidence of bunching on the left side of the cutoff in their work on Los Angeles restaurant inspections, as does Ho 2012 in San Diego and New York. But Figure 2.3 does raise two questions regarding the efficacy of the inspection process: first, does bunching represent the true distribution of restaurant cleanliness, or inspectors' unwillingness to give restaurants non-A grades?

[23]The small peaks at even numbers are due to the fact that there are no one-point violations, and many of the most common violations are worth two points. See Health Department 2016.

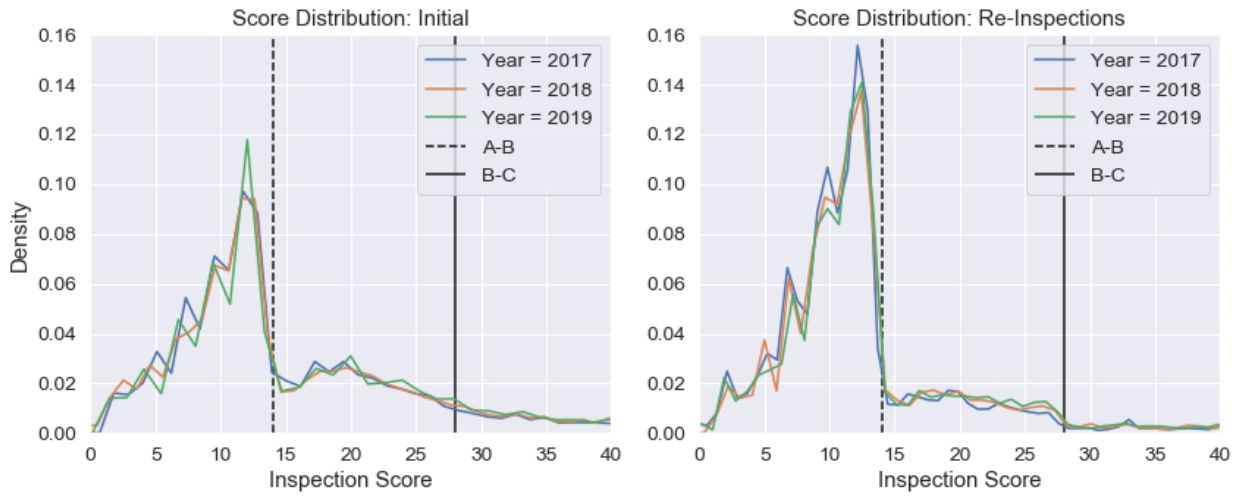


Figure 2.3: Distribution of Inspections Scores by Inspection Type and Year.

Description: Gaussian kernel density estimates (bandwidth = 0.5) of scores for initial inspections and re-inspections in the primary inspection sample.

Previous work, such as Dai and Luca 2020 suggests consumers take restaurant grades into account when making dining decisions. Biased signals of a restaurant's cleanliness reduces the efficiency of the consumer's decision-making process. It's not obvious from Figure 2.3 whether inspectors are hesitant to give non-A scores, since restaurants may invest in cleanliness in order to minimize costs while maximizing the probability of receiving an A grade. In such a setting, restaurants would naturally bunch just to the left of the cutoff.

The second question is whether so many restaurants pass on re-inspection because they use the initial inspection as a free pass and "clean up their act" when they know a re-inspection is shortly forthcoming. In such a situation, a restaurant's food poisoning risk is not reflective of its risk on inspection day, again leading consumers to use inaccurate information when making dining decisions.

I turn my attention to these two questions now. To lead with the conclusion, I find that inspectors appear to be hesitant to award non-A grades, and I do not find evidence that restaurants dynamically change their food safety profile over time.

2.4.2 Inspector Reluctance Towards Non-A Grades

In this section I develop reduced-form evidence of inspectors' local aversion to B grades. In particular, I take advantage of the fact that inspectors have some discretion in the number of points they assign for a given violation. The strategy is to condition on the set of observed violations, calculate the expected inspection score based on the severity of those violations, and compare that to the actual score. If inspections whose predicted scores are near the A-B boundary exhibit bunching, this suggests that inspectors are knocking a point or two off some violations in order to reach an A. This strategy is similar to that used by Makofske 2020a, who notes that health inspectors in Los Angeles are much more likely to assign low point totals on certain violations when an inspection is near the A-B margin.

To perform this test, I regress inspection scores on dummies for observed violations, and use this model to construct predicted scores for each inspection. I then condition on an inspection's predicted score (rounded to the nearest integer) and plot the distribution of realized inspection scores. Figure 2.4 shows the results of this analysis. When the expected score of an inspection is 12, for example, actual scores are approximately normally distributed around 12. However, when the expected score is 14, the distribution peaks at 13 and drops rapidly across the A-B boundary. More interestingly, when the expected score is 16 (a B), scores exhibit a bimodal distribution with a small peak around 13 and another peak around 17. Once the expected score reaches 18 there is no longer any bunching around the A-B cutoff, suggesting that inspectors are only willing to give the benefit of the doubt on marginal cases close to the A-B boundary.

The strongest competing explanation for this data is one of restaurant manipulation. Restaurants will of course try to present themselves in the best possible light when an inspector comes knocking, either by hiding violations or reducing the severity of existing violations. Since I condition on observed violations, there is no concern about hidden violations contaminating inference on inspector behavior. However, Figure 2.4 could be consistent with an inspector who "calls it like they see it" if restaurants on the A/B margin take action to minimize the severity of existing violations. This would require a fair amount of sophistication from restaurants: they would need

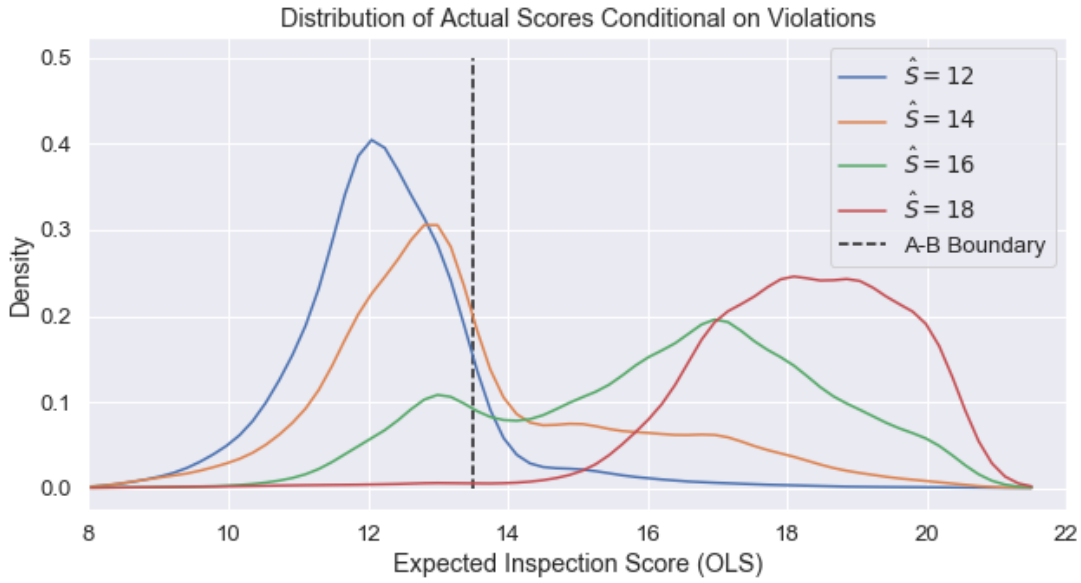


Figure 2.4: Score Distributions as a Function of Expected Score.

Description: Gaussian kernel density estimates (bandwidth = 0.5) of actual inspection scores conditioned on expected scores (based on the set of violations present).

to have sufficient knowledge of the health code to anticipate when they are on the A/B boundary, as well as how to minimize (but not eliminate) violations. However, there are two other reasons to be skeptical that restaurants are the driving force behind the bunching in Figure 2.4.

First, many violations are difficult for restaurants to manipulate in real-time. Some of the most frequent violations relate to features of the facility, such as its walls or plumbing, that are fixed for the duration of an inspection.[24] Other violations, particularly those related to pests, are highly random in nature. For example, conditional on filth flies being present, it would seem to be difficult for a restaurant to control whether an inspector sees 5 or 6 flies (the cutoff for a 5 or 6 point violation).[25] Of the 18 most frequent violations from 2011-16, 50% were structural or pest-related.

Second, and more importantly, violation-level scoring behavior is inconsistent with a restaurant

[24]From 2011-16, violations 10F ("Non-food contact surface improperly constructed."), 8A ("Facility not vermin proof"), 10B ("Plumbing not properly installed or maintained"), and 4A ("Food Protection Certificate not held by supervisor of food operations.") comprised 32% of all violations.

[25]From 2011-16, pest-related violations (4L - mice, 4N - filth flies, and 4M - roaches) comprised 11.8% of all violations.

mitigation story. If restaurants strategically reduce the severity of some violations to achieve A grades, average scores for manipulable violations should be lower in marginal inspections. Figure 2.5 plots the difference in average violation scores for marginal inspections (those with a predicted score between 13.5 and 17.5) and non-marginal inspections. Interestingly, pest-related violations tend to be scored less harshly in marginal inspections, whereas violations related to immediate food safety (improper sanitation or food temperatures) are penalized more harshly. My interpretation is that inspectors push marginal restaurants towards an A when some violations are random (such as seeing a bug), but push them towards a B if they see unsanitary behavior in the restaurant's control. This seems more plausible than restaurants being able to manipulate the exact number of cockroaches an inspector happens to see.

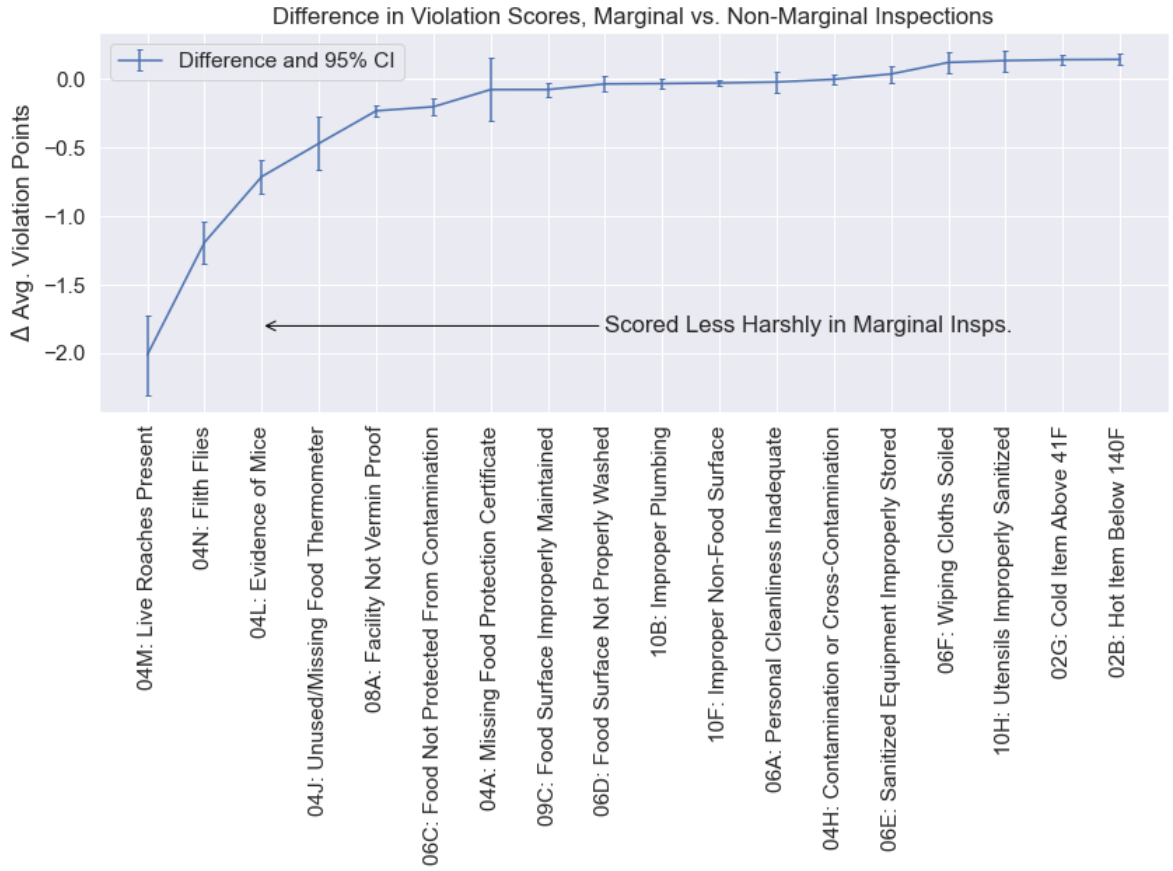
2.4.3 Restaurant Food Poisoning Risk Dynamics

To test whether restaurants "clean up their act" in response to initial inspections (or whether they vary their food safety more broadly over the inspection cycle), I construct a weekly dataset consisting of each restaurant's current inspection state (whether it is awaiting a re-inspection, and if not, its grade and inspection cycle timing); the time since their last inspection; the number of 311 food poisoning cases reported that week; and restaurant level controls (median inspection score quintile, cuisine, borough, and chain status). I then regress the number of 311 food poisoning reports against dummies for the inspection state, dummies for the number of months since their last inspection, and restaurant level controls.

Table 2.5 summarizes the results of this regression. Coefficients represent annual 311 food poisoning reports per 100 restaurants. Unsurprisingly the R^2 is nearly 0, since it's not feasible to predict in which particular week and restaurant a consumer will become sick. The excluded inspection state is an A grade restaurant on a 12 month inspection cycle (i.e. a restaurant that received an A on its initial inspection).

Encouragingly for the health department, inspection grades and scores are correlated with the underlying risk of a food poisoning report. Restaurants whose median inspection score is in the

Figure 2.5: Violation Scores for Marginal vs. Non-Marginal Inspections



Description: This figure uses 2011-2016 data to plot the difference in average violation scores for marginal inspections (predicted score between 13.5 and 17.5) and non-marginal inspections (all others). I plot $v_m - v_{nm}$, where $v_m = \mathbb{E}(s|\text{marginal, violation} = v)$, and v_{nm} the same for non-marginal inspections. Confidence intervals are calculated assuming $Var(v_m - v_{nm}) = Var(v_m) + Var(v_{nm})$.

Table 2.5: 311 FP Reports over the Inspection Cycle

	Effect of Inspection State on 311 FP Reports		
	(1)	(2)	(3)
Regulatory State (excluded = A-12)			
A-4	1.40**	0.96	0.95
	(0.68)	(0.68)	(0.69)
A-6	1.02**	0.78*	0.77*
	(0.42)	(0.42)	(0.42)
B-4	0.97	0.56	0.69
	(1.29)	(1.29)	(1.31)
B-6	2.21**	1.98**	2.12**
	(0.89)	(0.89)	(0.91)
C-4	3.26**	2.67	2.82*
	(1.65)	(1.65)	(1.64)
Pending Re-Inspection	1.78***	1.50**	1.51**
	(0.61)	(0.61)	(0.61)
Median Score Quintile (excluded = 1)			
2	0.12	-0.02	-0.02
	(0.44)	(0.44)	(0.44)
3	0.90*	0.87	0.87
	(0.53)	(0.54)	(0.54)
4	1.53***	1.48***	1.48***
	(0.42)	(0.43)	(0.43)
5 (Bottom)	1.79***	1.62***	1.60***
	(0.54)	(0.55)	(0.55)
Months Since Insp. (excluded = 0)			
1	-0.20	-0.26	-0.26
	(0.54)	(0.54)	(0.54)
2	-0.10	-0.18	-0.18
	(0.61)	(0.61)	(0.61)
3	0.41	0.31	0.32
	(0.61)	(0.61)	(0.62)
Future Results			
Upgrade within month			-2.73
			(2.71)
Downgrade within month			0.25
			(0.91)
Fixed Effects	None	Boro, Cuisine, Chain, Year	
R^2	0.000	0.001	0.001
N		1,902,114	

$^{*}p < 0.1,^{**}p < 0.05,^{***}p < 0.01$

Description: Unit of observation is a restaurant-week. Dependent variable = Weekly 311 FP reports*5200 (coefficients represent annual 311 FP reports per 100 restaurants). Sample includes 17,636 restaurants that received an initial or re-inspection in the primary dataset between 2017 and 2019. I only include restaurants weeks in which a restaurant was active. Regulatory states are denoted by their posted grade and the re-inspection cycle (e.g. A-4 means an A grade with a 4 month re-inspection cycle). Regressions include FE for up to 13 months following inspection, but only three are reported here; none of the coefficients are significant.

bottom quintile generate 1.5-2 more 311 food poisoning reports annually (per 100 restaurants). To give a concrete example, model (1) predicts that a group of 100 top-quintile, A-12 restaurants would generate 5.3 311 food poisoning reports annually, whereas a group of 100 bottom-quintile, C-4 restaurants would generate 10.3, nearly twice as many. Interestingly, restaurants that pass their re-inspection but received a B or C on their initial inspection seem to have 20-30% more food poisoning reports than A-12 restaurants, suggesting not all A grades are created equal.

Restaurants that are pending re-inspection do not appear to clean up their act — they generate food poisoning reports at a similar rate as restaurants that have recently undergone a re-inspection. In addition, the number of months since an inspection is not predictive of 311 food poisoning reports, suggesting restaurants are not easing back on sanitary processes during months when they won't be inspected and ramping back up when an inspection is forthcoming. There is certainly within-restaurant variation in food poisoning risk over time, but it does not appear to be systematically related to the inspection cycle.

There are several potential explanations for this lack of dynamics, which I won't attempt to disentangle here: (i) sanitation practices may require fixed investments (e.g. equipment) that requires little ongoing cost to maintain; (ii) sanitation practices may become ingrained as part of the "DNA" of a restaurant's day-to-day operations, such that consciously altering them would be an expensive complication; and (iii) restaurants and restaurant employees continue to face reputational and career risks for lax health practices even in the absence of formal inspections. The fixed investment story is supported by the data, as some of the most frequent violations relate to relatively fixed properties of the restaurant, such as the material or installation of non-food contact surfaces (10F); whether the building is vermin proof (8A); and whether plumbing is appropriately installed (10B). The model of Section 2.5 lets restaurants make an optimal one-time investment in their sanitation practices based on Health Department policies.

In model (3) I include indicators for whether a restaurant is upgraded or downgraded in the following month, with the intuition being that improved inspection scores may reflect previous

changes in health practices.[26] While the coefficients are the expected sign (restaurants that are about to be upgraded have fewer than expected food poisoning reports, and vice versa), the data is noisy, and I am unable to reject the null that future changes in state have any bearing on current risk.

2.4.4 Determinants of Inspector Bias

Section 2.4.2 suggests inspectors are biased towards A grades, perhaps informed by the makeup of inspection violations. This is consistent with my conversations with inspectors. They describe developing a feel for which restaurants are health risks. If the inspection "feels" like an A but the official score is a 14 or 15, the inspector might drop a minor violation or downgrade the severity of a violation in order to land on a 13. However, there may be other explanations. Cynical readers may suspect below-board behavior, with restaurants offering money or other compensation to inspectors that look the other way. Publicly reported examples of inspectors accepting bribes are rare and have serious consequences for the inspector, including possible jail time.[27]

The 2011-16 database, which includes anonymized inspector identifiers, can shed light on factors that contribute to inspector bias. I find four factors that predict increased inspector bias: (i) having many cases recently overturned by the OATH board; (ii) being later in their career; (iii) if a restaurant has previously been shut down by the Health Department; and (iv) the age of the restaurant (more A-bias at younger restaurants).

To begin, I define a proxy for inspector bias. Define s_{ir} as the random variable denoting the score inspector i would give restaurant r upon inspection. Let

$$\Delta_{ir} \equiv P(s_{ir} \in \{12, 13\}) - P(s_{ir} \in \{14, 15\})$$

be a measure of inspector bias towards A grades in restaurant r.[28] Since the same inspector rarely

[26] I define an upgrade as moving from a B or C grade to an A, and a downgrade as moving from an A to a B, C, or re-inspection.

[27] For example, see Investigation Department 2008, in which a health inspector was charged with accepting a $500 bribe and faced up to seven years in prison.

[28] I use two scores to the left and right of the A-B boundary because many violations have a minimum point score

inspects a restaurant more than once, it is impossible to measure Δ_{ir}. However, due to random assignment of inspectors to restaurants, the expected value of Δ_{ir} is identified at an inspector and a restaurant level. Let R_i denote the set of restaurants inspector i has visited, and I_r the set of inspectors that have visited restaurant r. Then:

$$\begin{aligned}
\Delta_i &\equiv \frac{1}{|R_i|} \sum_{r \in R_i} \mathbb{I}(s_{ir} \in \{12, 13\}) - \mathbb{I}(s_{ir} \in \{14, 15\}) \\
&\rightarrow_p \mathbb{E}_r(\Delta_{ir}) \\
\Delta_r &\equiv \frac{1}{|I_r|} \sum_{i \in I_r} \mathbb{I}(s_{ir} \in \{12, 13\}) - \mathbb{I}(s_{ir} \in \{14, 15\}) \\
&\rightarrow_p \mathbb{E}_i(\Delta_{ir})
\end{aligned}$$

Figure 2.6 shows that inspector bias is quite heterogeneous. For most inspectors Δ_i is estimated to lie between 0.05 and 0.15, but there is a long tail of inspectors who assign 12/13 scores 15-30 percentage points more often than 14/15 scores.

Table 2.6 shows that inspector bias increases with the number of inspections an inspector has recently had overturned by an OATH hearing. Since only inspections that generate a score of 14 or higher will ever be overturned, there is a mechanical negative correlation between Δ_i and the fraction of overturned cases. To correct for this, the dependent variable in Table 2.6 is the value of Δ_i when restricted to inspections that are not overturned.[29] The interpretation of Model (1) is: suppose an inspector has two quarters, one in which half of his inspections are overturned, and another in which none are overturned. In the first quarter the difference in probabilities between 12/13 and 14/15 scores for the un-overturned inspections will be about 6.3 percentage points larger than in the second quarter.

There are theoretical reasons why overturn rates might influence inspector decision making. Inspectors may be asked to testify as part of the OATH hearing, and in periods with many overturned

of 2, meaning it is mechanically difficult to reach certain scores.

[29]By random assignment, the distribution of scores in non-overturned inspections is independent of the number of overturned cases.

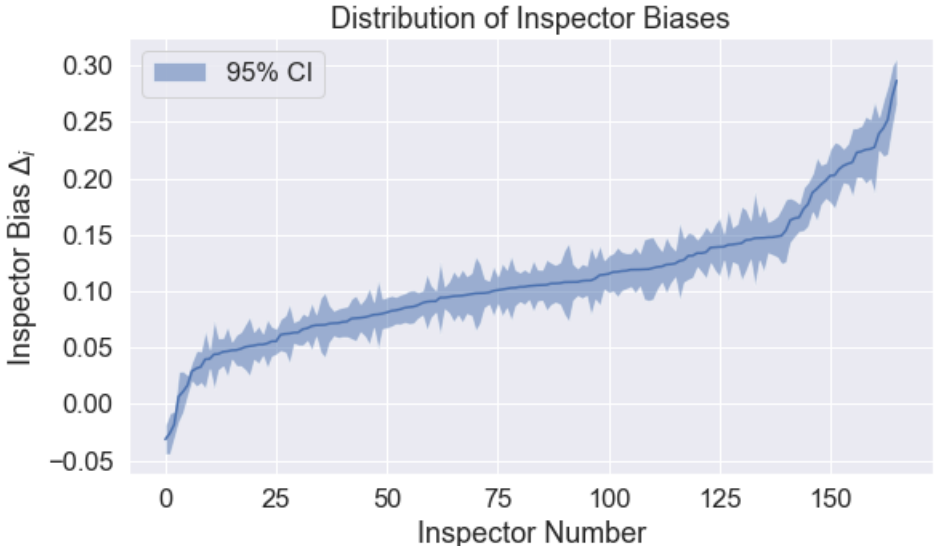


Figure 2.6: Estimated Distribution of Inspector Bias Δ_i.

Description: Fraction of inspections scored 12/13 - fraction of inspections scored 14-15 at an inspector level. Includes 166 inspectors that had at least 350 inspections in the 2011-16 dataset (covers 95% of all inspections). Scores are pre-OATH.

cases inspectors may err on the side of A grades in order to minimize further administrative burden. Second, having many inspections overturned may send a negative signal about the inspector's accuracy, again leading inspectors to err on the side of A's as their overturned caseload mounts. The one-sided nature of OATH hearings (no restaurant ever contests their A grade) is reminiscent of Leaver 2009, in which a threat of regulated entities publicizing mistakes leads some bureaucrats to take overly cautious actions.

Table 2.7 investigates restaurant-level bias Δ_r. Most notably, it shows that inspector bias appears to be decrease among restaurants that have ever been closed by the Health Department, holding fixed a restaurant's average score. In addition, bias towards A grades appears to be stronger in younger restaurants. Other than lower levels of bias in the Bronx, location and cuisine variables do not explain much of the variation in restaurant-level bias.

Table 2.6: Determinants of Inspector Bias I

	(1)	(2)	(3)	(4)
Overturned	0.127**	0.141**	0.144**	0.109*
	(0.047)	(0.048)	(0.048)	(0.044)
FirstYear	−0.037**	−0.018**	−0.018**	−0.016**
	(0.004)	(0.007)	(0.007)	(0.007)
Insp FE	X	X	X	X
Year/Quarter FE		X	X	X
Boro Composition			X	X
Cuisine Composition				X
No. Inspectors/Quarters		166/24		
N		2,201		
Adj. R^2	0.409	0.416	0.417	0.435
* $p < 0.05$, ** $p < 0.01$				

Description: Unit of observation: Inspector-Quarter. Dependent variable = fraction of un-overturned inspections scored 12/13 minus fraction scored 14/15 (-1 to 1 scale). Sample Definition: all inspectors with at least 350 inspections in the 2011-16 dataset (covers 95% of inspections). Variable definitions: "Overturned" = % of inspections given an initial score of 14 or above and a post-adjudication score of 13 or below; "First Year" = dummy variable indicating whether this is the first year an inspector is observed in the data; "Inspector FE" = indicator for each inspector (coefficients omitted); "Year and Quarter FE" = five year dummies (2011-2016) and four quarter dummies (1-4) (coefficients omitted); "Boro Composition" = % of inspections occurring in each borough of New York (coefficients omitted); "Cuisine Composition" = % of inspections occurring in the 10 most frequent cuisine categories (e.g. American, Chinese; coefficient omitted)

2.5 A Model of the Inspector-Restaurant Interaction

This section presents a two-stage model of the restaurant inspection industry, which consists of a planning phase and an inspection phase. During the inspection phase, an inspector receives a noisy signal of the restaurant's health state and issues a grade based on that signal; restaurant payoffs are a function of their grade.

In the planning phase, restaurants invest in their health practices in order to maximize their expected payoffs in the inspection phase (net of investment costs). The restaurant's problem is similar to an efficiency wage model with stochastic — and potentially biased — monitoring: restaurants trade off the effort of improving their health state against the lost wages associated with receiving

Table 2.7: Determinants of Inspector Bias II

	Coefficient	Standard Error
Intercept	0.018	0.019
Ever Closed	−0.058**	0.005
Chain	0.007	0.005
Boro (excluded = Bronx)		
Brooklyn	0.034**	0.006
Manhattan	0.024**	0.006
Queens	0.035**	0.006
Staten Island	0.032**	0.010
Cuisine Category (excluded = African)		
American	0.019	0.018
Asian	0.007	0.019
Central/South American	-0.006	0.019
Drinks	0.007	0.023
European	0.007	0.018
Middle Eastern	0.008	0.026
Other	0.004	0.019
Sweets	0.011	0.019
Mean Score Quantile (excluded = 1)		
2	0.070**	0.004
3	0.133**	0.005
4	0.322**	0.005
5 (bottom)	0.080**	0.004
Year of First Inspection (excluded = 2016)		
2017	-0.001	0.004
2018	0.027**	0.005
2019	0.045**	0.008
2020	0.035*	0.016
N	17,095	
Adj. R^2	0.262	
* $p < 0.05$, ** $p < 0.01$		

Description: Unit of observation = Restaurant. Dependent variable = fraction of observations scored 12/13 minus fraction score 14/15 (-1 to 1 scale). Sample Definition: all restaurants with cuisine data and at least one inspection or re-inspection after 2016 in the primary dataset. "Ever Closed" indicates a restaurant was ever closed by the DOHMH between 2016-2020.

a poor grade.

Restaurant Health States During the planning phase, restaurants choose a health state $x \in [0, \bar{X}]$, where, in keeping with NYC's scoring conventions, higher values of x denote worse health practices. The cost for restaurant r to choose health state x is

$$c_r(x) = C_r \cdot (\bar{X} - x)$$

The parameter C_r measures the cost of improving the restaurant's health state by one unit. The upper bound $\bar{X}$ represents the fact restaurants cannot operate above a certain level of dis-hygiene. This could be because they are shut down by the health department or because conditions become so apparently dirty that customers stop buying.

Inspection Scores During the inspection phase, an inspector visits the restaurant and issues a score s. Based on this score, the restaurant is given a grade g and a time until its next inspection. Scores are generated according to an ordered probit with additive heteroskedasticity: an inspector reports score s if

$$\beta x + \epsilon \in (\theta_{s-1}, \theta_s), \quad \epsilon | x \sim \mathcal{N}(0, a + bx) \tag{2.1}$$

where θ_s are score-specific thresholds and $a \geq 0$ and $b > 0$ are variance parameters.[30] Under the assumption that β is non-zero, Appendix B.1 shows that there is no loss in generality by setting $(a, b, \beta) = (0, 1, 1)$.

The interpretation of the score-generating model is as follows: an inspector receives a noisy signal of the restaurant's true state x, and gives the restaurant a score associated with whatever bin that signal falls in. The noise term ϵ is drawn independently for each inspection, and is intended to capture inspection-specific shocks. These shocks could be health-related (for example, whether a bug happens to walk across the counter) or could capture random variation in the inspector's preferences from day-to-day. Unlike many applications of ordered probit models, I am particu-

[30] Assuming $b \neq 0$ means the model does not nest the standard ordered probit $(a, b) = (1, 0)$.

larly interested in the shape of the thresholds θ_s as they contain information about how sensitive inspectors are to changes in x at different points of the score distribution.

There are two homogeneity assumptions implicit in the ordered probit model:

- Thresholds θ_s are constant across inspectors.
- Inspection variance is constant across inspectors (no difference in inspector skill).[31]

While Section 2.4.4 suggests inspectors are heterogeneous in their scoring behavior, these assumptions have limited impact on estimates and counterfactuals. Due to random assignment of inspectors to restaurants, Appendix B.4 shows that in the presence of inspector heterogeneity, the model recovers – to first order – the mean inspector thresholds. This argument easily extends to heterogeneous variances as well.

Grading and timing conventions closely follow those of New York (see Figure 2.1). For restaurants that do not receive an A on initial inspection, re-inspections happen immediately with an independent draw of ϵ.

The Restaurant's Problem I allow restaurant payoffs to vary with their current grade.[32] Let π_{rg} denote the payoff restaurant r receives from having grade A. Similarly, let $s_g(x)$ denote the fraction of time a restaurant has grade g given health state x. Appendix B.2 gives an explicit expression for $s_g(x)$.

The restaurant's problem is

$$\max_{x \in [0, \bar{X}]} s_A(x)\pi_{rA} + s_B(x)\pi_{rB} + s_C(x)\pi_{rC} - c_r(x)$$

Define $\Delta_{rg} \equiv \pi_{rg} - \pi_{rC}$, the loss in restaurant payoff when moving from grade g to C. With this notation and the fact that $s_A(x) + s_B(x) + s_C(x) = 1$, the restaurant's first order condition can

[31] For recent work that investigates the impact of heterogeneous skill on decision making, see Chan et al. 2019.

[32] The rationale for only allowing payoff to vary by the public grade, and not the inspection cycle, is that consumers only observe the grade.

be written

$$\frac{C_r}{\Delta_{rA}} = -\left[s'_A(x) + s'_B(x)\frac{\Delta_{rB}}{\Delta_{rA}}\right]$$

which says that the cost of improving the health state by one unit, relative to the gains from an A grade, must be weighed by the expected change in the grade distribution.[33] Let $\tilde{C}_r = \frac{C_r}{\Delta_{rA}}$ be a restaurant's normalized costs and $\tilde{\rho}_r = \frac{\Delta_{rB}}{\Delta_{rA}}$ the ratio of B-C losses to A-C losses. The first-order condition is then:

$$\tilde{C}_r = -\left[s'_A(x) + s'_B(x)\tilde{\rho}_r\right], \tag{2.2}$$

which is useful for identification. I assume:

Assumption 2.5.1. *The restaurant's payoffs satisfy* $\pi_{rA} \geq \pi_{rB} \geq \pi_{rC}$.

Assumption 2.5.1 implies $\tilde{\rho}_r \in [0, 1]$; a ratio of 0 means there is no loss associated with moving from a B to a C, while a ratio of 1 means there is no loss associated with moving from an A to a B.

2.5.1 Identification

The model consists of parameters $\Theta = \{\theta_s, C_r, \pi_{rg}\}$ and data $\mathcal{D} = \{P_r(s)\}$, where $P_r(s)$ represents the cumulative distribution of scores at restaurant r. Another useful object, which is neither a primitive nor data, is the set of restaurant latent states $x_r(\Theta)$. In this section I summarize identification results given the following assumption:

Assumption 2.5.2. *There exist at least two restaurants whose score distributions are distinct and have full support.*

Assumption 2.5.2 is relatively weak. By the normality assumption (Equation 2.1), the score distribution for any restaurant with a latent state $x \neq 0$ will have full support, so Assumption 2.5.2 rules two cases: (i) the case where every restaurant receives perfect scores on every inspection; and (ii) the case where every non-perfect restaurant has the exact same distribution of scores.

[33] It's possible for the FOC to fail, in which case $x = 0$ or $\bar{X}$.

Proposition 2.5.1. *Under Assumption 2.5.2, inspector thresholds θ_s and restaurant latent states $x_r(\Theta)$ are identified. If $x_r \in (0, \bar{X})$, then, for any $\tilde{\rho}_r$, restaurant r's normalized costs $\tilde{C}_r(\tilde{\rho}_r)$ is identified as well.*

This proposition says that the best the current data can do is identify *normalized* costs as a function of a set-identified parameter $\tilde{\rho}_r$, which turns out to be informative enough for counterfactual purposes. Separately identifying C_r and π_{rg} would require restaurant-level profit data across different grades. Assumption 2.5.1 is not required for identification.[34]

Proof. Let $\mathcal{R}^f$ denote the set of restaurants whose score distribution has full-support. For $r \notin \mathcal{R}^f$, it must be the case that $x_r = 0$, so the latent state for those restaurants are known.

For any $r \in \mathcal{R}^f$

$$P_r(s^*) \equiv P(s \leq s^*|x_r) = \Phi\left(\frac{\theta_{s^*} - x_r}{\sqrt{x_r}}\right)$$

Defining $d_{rs} \equiv \Phi^{-1}(P_r(s))$, which is observable, and inverting gives

$$\theta_s = x_r + \sqrt{x_r} d_{rs} \tag{2.3}$$

Therefore given x_r for any $r \in \mathcal{R}^f$, the inspector thresholds θ_s are identified.

Moreover, differencing Equation 2.3 for $r, r' \in \mathcal{R}^f$ gives

$$x_{r'} + \sqrt{x_{r'}} d_{r's} = x_r + \sqrt{x_r} d_{rs} \tag{2.4}$$

Thus if x_r is known for one restaurant, it is known for every restaurant in $\mathcal{R}^f$.[35]

To show that x_r is identified for some r, let restaurants 1 and 2 denote the two restaurants with distinct, full-support score distributions. WLOG, suppose $d_{12} - d_{11} \neq d_{22} - d_{21}$. For $r \in \{1, 2\}$ and

[34] Assumption 2.5.1 is useful for empirical work since it puts bounds on the cost parameter $\tilde{C}_r$.

[35] From the quadratic formula:

$$\sqrt{x_{r'}} = \frac{-d_{r's} + \sqrt{d_{r's}^2 + 4(x_r + \sqrt{x_r} d_{rs})}}{2}$$

where I take this positive solution since $\sqrt{x_{r'}} > 0$.

$s \in \{1, 2\}$, Equation 2.3 generates a system of four equations in the four unknowns $(\theta_1, \theta_2, x_1, x_2)$. Let $c \equiv \frac{d_{12}-d_{11}}{d_{22}-d_{21}}$, which is well-defined by the full-support assumption. Solving gives

$$x_1 = \left(\frac{d_{11} - d_{21}c}{c^2 - 1} \right)^2$$

Since $c \neq 1$, x_1 is well-defined, and, by the arguments above x_r and θ_s are identified.

Lastly, if $x_r \neq 0$ or $\bar{X}$, restaurant r's normalized costs are identified as a function of $\tilde{\rho}_r$ through the first-order condition Equation 2.2. □

2.6 Estimation and Results

2.6.1 Maximum Likelihood Estimation

The biggest challenge to identification is recovering each restaurant's chosen state, x. The easiest way to do this is via maximum likelihood: the probability that a restaurant with health state x receives a score s is

$$\begin{aligned} P(s|x) &= P(x + \epsilon \in (\theta_{s-1}, \theta_s)|x) \\ &= P\left(\frac{\epsilon}{\sqrt{x}} \in \left(\frac{\theta_{s-1} - x}{\sqrt{x}}, \frac{\theta_s - x}{\sqrt{x}} \right) \middle| x \right) \\ &= \Phi\left(\frac{\theta_s - x}{\sqrt{x}} \right) - \Phi\left(\frac{\theta_{s-1} - x}{\sqrt{x}} \right) \qquad (2.5) \\ &\approx \frac{\theta_s - \theta_{s-1}}{\sqrt{x}} \varphi\left(\frac{\theta_s - x}{\sqrt{x}} \right) \qquad (2.6) \end{aligned}$$

The full likelihood for the sample can then be written

$$\mathcal{L}(\theta, x) = \prod_r \prod_i P(s_i|x_r)$$

where r indexes restaurants and i inspections.[36]

[36] In the main results I pool initial inspections and re-inspections. Separating these inspections has little impact on the conclusions: the correlation between restaurant health states and inspector preferences when estimated on disaggregated data are 99.5% and 93.7%, respectively. This analysis is available upon request.

For a given vector of thresholds θ, each restaurant's x only affects the likelihood of their own inspection results. Thus one approach to estimation would be an outer routine that maximizes over θ, and an inner routine that maximizes Equation 2.5 for a fixed θ. However, given that there are tens of thousands of restaurants in the sample, estimating restaurant-level latent states without a closed-form expression is computationally burdensome. To alleviate this burden, I split restaurants into 50 groups based on their average inspection score. This grouping is motivated by the observation that the score distribution increases in an FOSD sense with x, meaning the expected inspection scores are an increasing function of x as well.[37]

The first-order approximation of the likelihood (Equation 2.6) admits a closed-form solution for x_r given θ:

$$x_r^*(\theta) = \frac{-1 + \sqrt{1 + \frac{4}{N_r} \sum_i \theta_{s_i}^2}}{2} \approx RMS(\theta_{s_i})$$

which is a useful starting point for the maximization routine.[38]

Identification Intuition The algebraic identification argument of Section 2.5.1 nicely outlines the patterns in the data that identify the restaurant's health state x and the inspector's thresholds θ. For simplicity, suppose the latent state x_0 is known for some restaurant.

Restaurant health states are determined by comparing their score distributions to the known benchmark. Substituting x_0 into the righthand side of Equation 2.4 and implicitly differentiating the lefthand side with respect to $d_{r's}$ gives:

$$\frac{dx_{r'}}{dd_{r's}} = -\frac{x_{r'}}{\sqrt{x_{r'}} + \frac{1}{2} d_{r's}} < 0$$

If $d_{r's} > d_{rs}$ for all scores s — if restaurant r' is always more likely than r to receive an inspection

[37] The score CDF is:

$$P(s \leq s^* | x) = \Phi\left(\frac{\theta_s}{\sqrt{x}} - \sqrt{x}\right)$$

whose inner argument is decreasing for any $\theta_s, x > 0$. The fact that $\theta_s > 0$ for all s is equivalent to saying that restaurants with $x = 0$ receive the lowest possible score.

[38] Alternatively, this approximate solution could be used to estimate restaurant-specific x's without the need for grouping.

below a given score — then $x_{r'} < x_r$, meaning restaurant r' has better health practices than r.

Inspector thresholds are determined by the distribution of scores at a particular restaurant. Differencing Equation 2.3 for scores s and $s-1$ gives:

$$\theta_s - \theta_{s-1} = \sqrt{x_{rs}}(d_{rs} - d_{r,s-1})$$

If a restaurant has very few inspections at score s, $d_{rs} - d_{r,s-1} \approx 0$, which implies $\theta_s \approx \theta_{s-1}$.

Another way to think about the identification of θ_s is in terms of residual inspection risk. The assumption of normally distributed inspection errors implies that, while restaurants may be able to control the mean of their inspection score, they cannot control the distribution of scores around that mean. This reflects the fact that some elements of inspections are out of a restaurant's control, such as the date and time an inspector shows up, the inspector's mood, and randomness of a given inspection (such as a mouse sighting). The model interprets sharp changes in a restaurant's score distribution as evidence of kinks in the inspector's preferences θ_s, since it does not believe restaurants could manipulate scores so precisely on their own.

2.6.2 Results

Figure 2.7 shows the estimated values of the inspector thresholds θ_s and restaurant's latent health states x_r (lower x_r is better). The right-hand side shows that approximately 80% of restaurants have chosen a health state that will earn an A grade on more than half their inspections. Very few restaurants have chosen a health state that would consistently earn a C grade, which is not too surprising as those restaurants would soon be shut down by the Health Department.

Inspector thresholds exhibit a noticeable kink around the A-B boundary. One useful way to interpret this graph is through the inspector's optimal score function, $s^*(x) \equiv \min\{s|\theta_s \geq x\}$, which satisfies

$$s^*(\theta_s) = s$$

The left-hand side of Figure 2.7, therefore, can be read with the y-axis as the latent state x and the

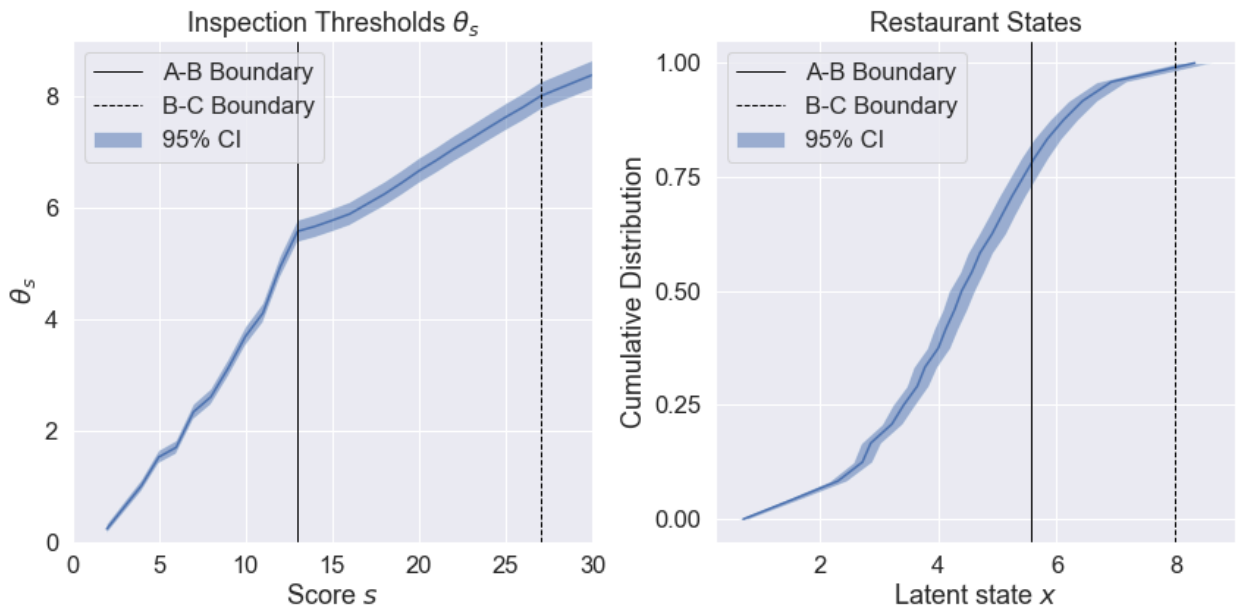


Figure 2.7: Estimated Inspector Thresholds and Restaurant States

Description: Confidence intervals based on plug-in estimate of MLE asymptotic variance.

x-axis as the optimal score $s^*(x)$. Below the A-B boundary, a one point increase in x generates approximately a 2-point increase in the score, whereas above the A-B boundary, a one point increase in x generates a 5-point increase in the score.

Interpreting the shape of the inspector thresholds depends on the goal of inspection scores. One natural interpretation is that scores should reflect underlying food poisoning risk. In this case, a kink like in Figure 2.7 could be justified if there is a sudden increase in food-poisoning risk around the A-B boundary. To pursue this idea, I estimate a flexible model of food-poisoning risk as a function of the latent state x:

$$FP_{rt} = \beta_0 + \beta_x x_r + \sum_{d=2}^{10} \beta_d D_{rd} + \beta_c C_r + \epsilon_{rt} \tag{2.7}$$

where FP_{rt} are the number of food poisoning complaints submitted to the city's 311 system at restaurant r in month t; D_{rd} is an indicator if x_r is in the d-th decile of the x distribution; and C_r is a set of restaurant controls, namely their borough, cuisine type, and chain status.

Figure 2.8 shows the estimates of food-poisoning risk. The left side suggests, perhaps unsurprisingly, that excess food-poisoning risk is concentrated in the lower half of the restaurant distribution, with most of the change in food poisoning risk occurs between $x = 4$ and $x = 6$ (the 37-86th percentiles). Given the pronounced S-shape of the food poisoning risk curve, I construct a smoothed logistic estimator

$$\widehat{FP}(x) = FP_0 + \frac{c}{1 + \exp\left(-k(x - x_0)\right)} \tag{2.8}$$

where the parameters $\{FP_0, c, k, x_0\}$ are fit to minimized the average squared distance between $\widehat{FP}(x)$ and the OLS estimates from Equation 2.7.

The right-hand side of Figure 2.8 shows a parametric plot of $s^*(x)$ against $\widehat{FP}(x)$, along with corresponding quantiles of the x distribution. To the Health Department's credit, restaurants with the most predicted food poisoning cases receive the highest scores. However, it is odd how little the variation in scores is related to variation in FP risk. Much of the variation in scores comes from the bottom half and the top quartile of the x distribution, where there is very little change in the risk of food-poisoning. Conversely, most of the change in food-poisoning risk occurs between the 40-75th percentiles of the x distribution, where restaurants receive scores of 12-13 (an A).

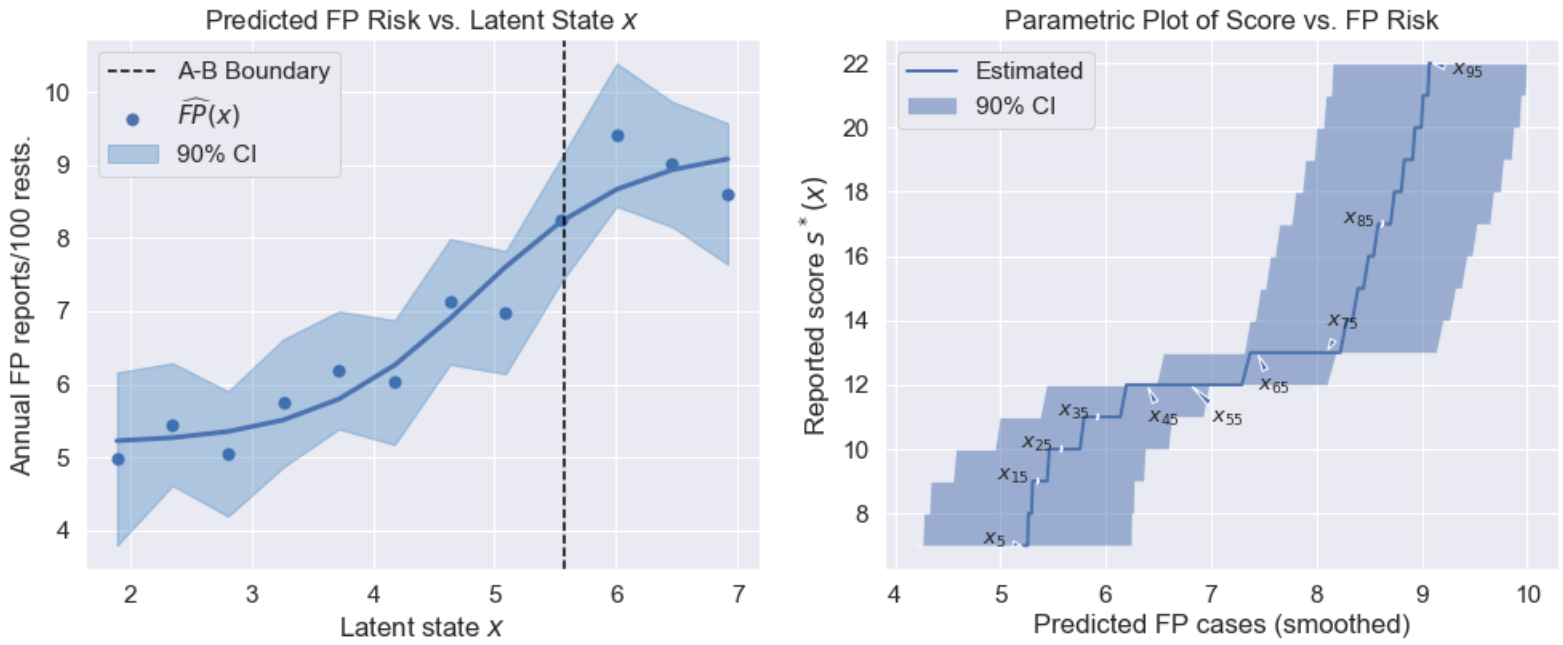


Figure 2.8: Inspection Scores, Food Poisoning Risk, and the Restaurant's Latent State.

Description: (Left hand side) Predicted annual food poisoning cases (per 100 restaurants) and 90% CI based on OLS estimates of Equation 2.7. Estimates are for an American, non-chain restaurant in Manhattan. Smoothed predictor is based on fitting a logistic curve $\widehat{FP}(x) = FP_0 + \frac{c}{1+\exp(-k(x-x_0))}$ to OLS estimates. (Right hand side) Parametric plot of $s^*(x)$ against $\widehat{FP}(x)$. Quantiles of the x distribution are marked by x_q. Confidence interval reflects the values of $\widehat{FP}(x)$ when fit using the upper and lower confidence bounds of the OLS estimates.

2.7 Counterfactuals

Motivated by Figure 2.8, in this Section I estimate the response of food poisoning totals to changes in the A-B boundary. In addition, I consider the effect of tweaking two institutional details: (i) removing re-inspections; and (ii) removing preferential re-inspection timing for restaurants that receive A grades. Table 2.8 summarizes the key outcomes for the three counterfactuals, which I describe in greater detail below.

2.7.1 Procedure and Assumptions

For each counterfactual, I first re-calculate each restaurant's optimal health state x using the first-order condition Equation 2.2. I do this for a grid of values of $\tilde{\rho} \in [0, 1]$. Then, using these updated health states, I estimate food-poisoning levels using the smoothed estimator from Equation 2.7.

The key assumption maintained in these counterfactuals is that consumer demand as a function of restaurant grade is constant. This is important for two reasons. First, if consumer demand changes, both the left and right-hand side of the first-order condition change, as $\tilde{C}_r$ and $\tilde{\rho}_r$ both contain grade-specific payoffs π_{rg}. In addition, if either aggregate demand or relative demand by grade changes, the predictive model (Equation 2.7) would no longer be valid.

Assuming constant consumer demand would be valid if, for example, reforms are largely invisible to consumers. Consumers continue to see A/B/C grades and interpret them as they did previously. However, if the information content of a particular grade changes substantially, consumers may learn this and change their eating patterns. Given that each counterfactual scenario induces little change in the prevalence of A, B, and C grades, or in the relative food-poisoning risk across grades, I would be surprised if changes in consumer demand swamped the effects shown in the counterfactuals.

Whether constant consumer demand is a conservative assumption or not depends on how diners shift their demand. Let (w_g, f_g) represent the fraction of meals consumed and the number of food-

poisonings per meal at grade g restaurants. The change in the aggregate food-poisoning rate can be decomposed as:

$$\sum w_g^c f_g^c - \sum_g w_g f_g = \underbrace{\sum_g w_g (f_g^c - f_g)}_{\text{A: change in rates}} + \underbrace{\sum_g (w_g^c - w_g) f_g^c}_{\text{B: change in demand}} + \underbrace{\sum_g (w_g^c - w_g)(f_g^c - f_g)}_{\text{C: interaction}}$$

By leaving weights fixed, I calculate A in the above decomposition. If the change in weights is positively correlated with the reduction in food poisoning, C will be negative. In addition, if the change in weights is negatively correlated with counterfactual food poisoning risk, B will be negative as will. Therefore, under the assumption that demand shifts towards restaurants whose food-poisoning risk declines the most:[39]

- If clean restaurants improve the most, assuming constant demand underestimates improvements in the aggregate food-poisoning risk per meal
- If dirty restaurants improve the most, the effect of assuming constant demand on estimated aggregate food-poisoning risk per meal is ambiguous.

Another margin which the counterfactuals implicitly shut down is that consumers may substitute away from home-cooked meals in response to changes in the regulatory regime. Depending on the relative risk of home-cooked versus restaurant-cooked meals, this may lead to more of fewer cases of food poisoning. Jin and Leslie 2003 consider this question in the context of Los Angeles's introduction of grade cards, and find that food-poisoning risk for A-restaurant meals is not statistically different from that of home-cooked meals. Given the prevalence of A-restaurants in New York, this suggests that the home-to-restaurant margin is not a primary driver of aggregate food-poisoning risk.

[39] That is, that $\text{cov}(w_g^c - w_g, f_g^c - f_g) < 0$

Lastly, the fact that restaurant payoffs are fixed across counterfactuals has an interesting implication for policy design. Recall the restaurant's first-order condition, Equation 2.2:

$$\tilde{C}_r = -\left[s'_A(x) + s'_B(x)\tilde{\rho}_r \right]$$

For a fixed level of $\tilde{C}_r$, the optimal level of x decreases as the right-hand side ("RHS") increases. Integrating the right-hand side gives:

$$-\int_0^{\bar{X}} \left(s'_A(x) + s'_B(x)\tilde{\rho}_r \right) dx = s_A(0) - s_A(\bar{X}) + \tilde{\rho}_r \left(s_B(0) - s_B(\bar{X}) \right)$$

So long as a policy does not change grade shares at the extremes (e.g. if 0's receive A's and $\bar{X}'s$ receive C's), the average value of RHS is fixed across counterfactuals. This highlights an inherent policy tradeoff: improvements in health states (generated by increasing RHS for certain values of x) will be, at least in part, offset by declines in health states at values of x where RHS decreases. Given the non-linear nature of food-poisoning risk shown in Figure 2.8, effective policies will generate improvement for the restaurants where the food-poisoning risk curve is steepest.

Table 2.8: Summary of Counterfactual Outcomes

		Range of Outcomes		
	Baseline	Lower A-B Boundary	No Re-Inspections	No Preferential Timing
311 FP Reports/Baseline	1.00	(0.78, 0.87)	(0.86, 0.90)	(0.87, 0.90)
(Δ from Baseline)		(-0.22, -0.13)	(-0.14, -0.10)	(-0.13, -0.10)
Posted Grade Shares				
A	0.91	(0.80, 0.92)	(0.86, 0.88)	(0.92, 0.95)
		(-0.11, 0.01)	(-0.05, -0.03)	(0.01, 0.04)
B	0.06	(0.06, 0.18)	(0.08, 0.10)	(0.04, 0.06)
		(0.00, 0.11)	(0.02, 0.04)	(-0.02, 0.00)
C	0.03	(0.02, 0.03)	(0.03, 0.04)	(0.01, 0.02)
		(-0.01, 0.00)	(0.00, 0.01)	(-0.02, -0.01)
FP Risk by Grade				
$\mathbb{E}(FP\|A)$	6.71	(5.31, 5.79)	(5.74, 6.06)	(5.98, 6.16)
		(-1.40, -0.91)	(-0.97, -0.64)	(-0.73, -0.55)
$\mathbb{E}(FP\|B)$	8.11	(6.48, 7.13)	(7.26, 7.43)	(6.64, 7.30)
		(-1.63, -0.98)	(-0.84, -0.68)	(-1.46, -0.81)
$\mathbb{E}(FP\|C)$	8.67	(8.09, 9.19)	(8.31, 8.82)	(6.87, 7.80)
		(-0.59, 0.51)	(-0.36, 0.15)	(-1.80, -0.87)

Description: Key outcomes under the three counterfactuals: (i) lowering the A-B boundary; (ii) removing re-inspections; and (iii) eliminating preferential timing for strong performers (e.g. a 12 month cycle for A restaurants vs. a 4 month cycle for C restaurants). "Posted Grade Shares" means the average percent of the time a restaurant will have a given grade, equivalent to a point-in-time snapshot of the grade distribution. Ranges represent minimum and maximum estimated outcomes for $\tilde{\rho} \in [0, 1.0.9]$.

2.7.2 Changing Inspector Thresholds

The first question I investigate is whether the scoring policy shown in the righthand side of Figure 2.8 is driven by inspector behavior, or could be a rational policy chosen by a benevolent planner. Although the flat section of the scoring policy is indicative of inspectors preferring to give A grades, inspection scores and public grades have different functions. Grades are meant to inform the public about the relative health risks across restaurants. Scores, while available online, are much less prominent, and are designed to give inspectors a framework to judge health risks. Even with no distortions due to inspector preferences, we may not expect a perfectly linear relationship between scores and underlying health risks.

I therefore focus on a related question that abstracts from the precise score policy: would a benevolent planner locate the A-B boundary at the point indicated in Figure 2.8? The odd feature of this location is that the marginal risk of food-poisoning is highest near the boundary, suggesting large returns to improved health practices for restaurants near the boundary. Indeed, as shown in Figure 2.9, I estimate food poisoning cases would drop 13-22% if the A-B boundary were located between at 3.5 versus its current value of 5.6.[40]

However, a planner may also care about restaurant compliance costs in addition to aggregate food poisoning cases. Consider a planner whose utility from an A-B boundary θ is:

$$u(\theta) = -\left[\alpha \underbrace{\sum_r FP(x_r(\theta))}_{\text{aggregate FP cases}} + (1-\alpha) \underbrace{\sum_r \tilde{C}_r(\bar{X} - x_r(\theta))}_{\text{aggregate investment costs}}\right]$$

where α is the weight placed on aggregate food poisoning cases, and $1-\alpha$ the weight on restaurant effort. Note that the planner measures $\tilde{C}_r$, which, from Equation 2.2, is the ratio of effort costs to the lost value between an A and C grade. This implies a planner that is, all else equal, less

[40]Food poisoning is not a monotonic function of the location of the A-B boundary. If the boundary is set too low, restaurants are discouraged by the investment required to meet the impossibly high standards, and begin to worsen their health practices.

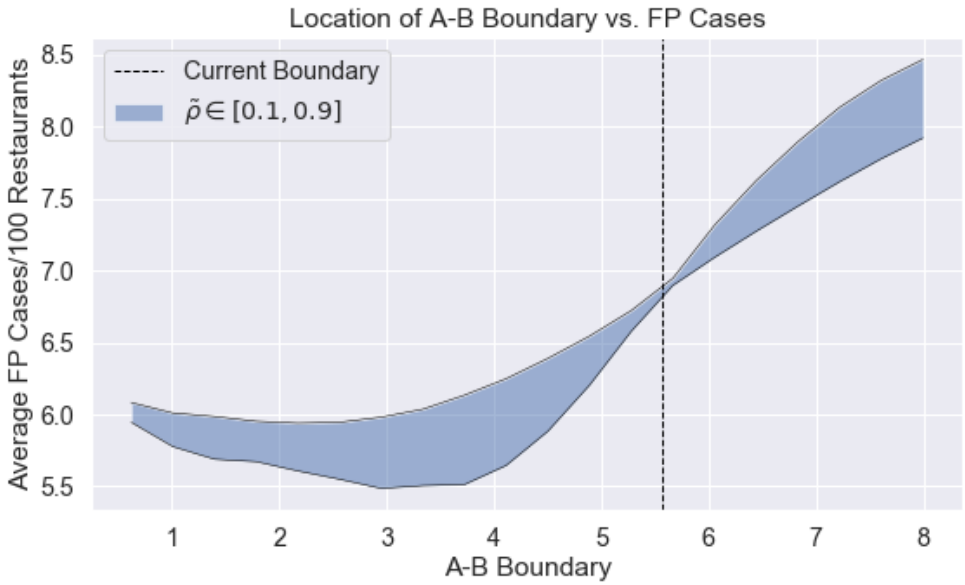


Figure 2.9: Aggregate Food Poisoning vs. A-B Boundary

concerned about effort costs for restaurants that receive large payoffs for compliance. Another reasonable objective would use effort costs C_r, but without further data it is not feasible to separate grade-specific payoffs from effort costs.

Figure 2.10 plots the planner's utility as a function of α and θ under the assumption that $\tilde{\rho}_r = 0.5$ (plots for other values of $\tilde{\rho}$ produce similar results). The planner's utility function is noticeably saddle-shaped. In the northwest corner, the planner puts low utility weight on food poisoning cases, and the optimal policy is to set a high threshold that restaurants can pass with minimal effort. As the utility weight on food poisoning increases, the optimal policy jumps to the "ridge" in the southeast corner. At the extreme of $\alpha = 1$, the planner's objective is to minimize total food poisoning cases, which, as indicated by Figure 2.9, happens when the boundary is a little below 3.

It is difficult to justify the current location of the A-B boundary as an optimal choice by a planner weighing food poisoning cases and compliance costs. Given the current boundary, the middle third of the restaurant distribution is clustered in an area where the marginal food poisoning risk is high, meaning small changes in the threshold can lead to large improvements in aggregate risk. This of course does not imply that *no* preferences can rationalize the observed inspection, but

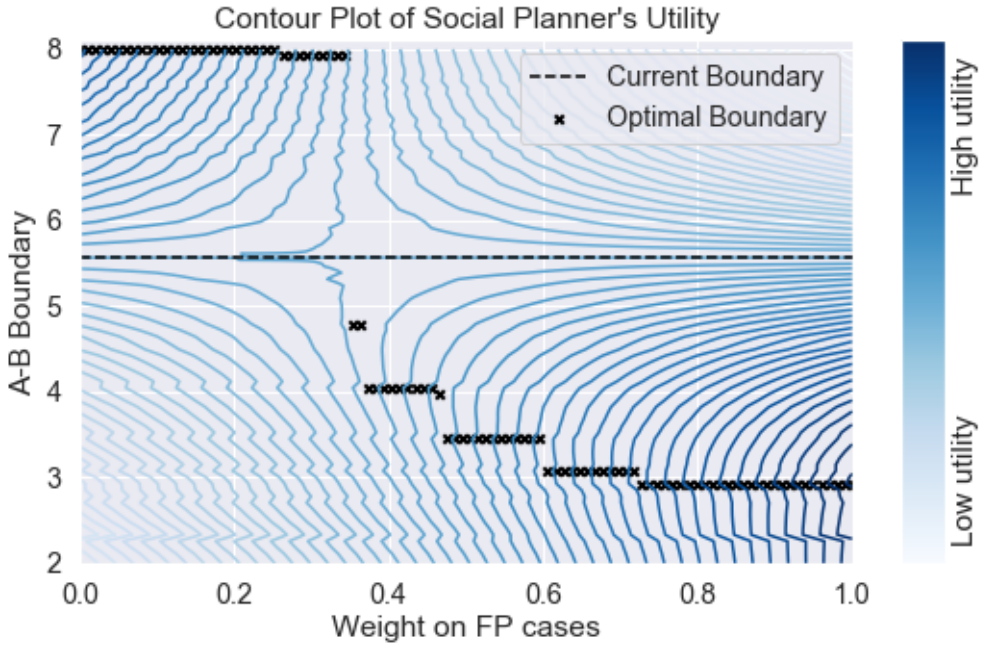


Figure 2.10: Planner's Utility for different α and A-B Boundaries

Description: For each boundary, counterfactual food poisoning cases are calculated using the approach outlined in Section 2.7.1. To make the food poisoning and cost scales similar, food poisoning is calculated as average 311 reports per 10 restaurants, as opposed to per 100 restaurants in the rest of the chapter. In this chart, $\tilde{\rho} = 0.5$; other values of $\tilde{\rho}$ produce similar figures.

rather that they would need to include additional criteria, such as the cost of inspections.

In Table 2.8, I report outcomes when the A-B boundary is set at 3.5, which corresponds to the value chosen by a planner with utility weight $\alpha = 0.6$. In this scenario, aggregate food-poisoning cases drop by 12.7-21.7%, with the largest improvements coming from B and C restaurants. Grade shares remain approximately the same, although there is a slight increase in the number of B and C restaurants in this scenario. Figure 2.11 shows how the restaurant's first-order condition changes in this scenario. For most restaurants, stricter inspections increase the marginal benefit of investing in their health state, as shown in Figure 2.11. However, for the bottom 0-7% of restaurants, the opposite effect occurs.[41] For these restaurants, the degree of improvement required to maintain good grades on stricter inspections becomes prohibitively expensive, leading them to worsen their

[41]The exact set of restaurants for whom this discouragement effect holds is sensitive to the value of $\tilde{\rho}$: fewer restaurants worsen their health states if $\tilde{\rho}$ is high, since restaurants are incentivized to maintain a B grade over a C.

health states.

2.7.3 Removing Re-Inspections

In this counterfactual, inspector thresholds are set at their original levels, but the inspection process is altered to remove re-inspections. Mechanically, this influences a restaurant's optimal choice of their health state x through the first-order condition, Equation 2.2.

Re-inspections give restaurants a second "at-bat" to receive a good grade. This has the effect of decreasing the marginal value of x for clean restaurants, and vice-versa. This is easiest to see in a simplified setting: suppose a restaurant can receive an A or B grade, and that an inspection generates an A grade with probability $P_I(A|x)$. Let $s_1(x)$ and $s_2(x)$ denote the probability a restaurant has an A grade with re-inspections and no re-inspections, respectively. Then:

$$s_1(x) = \underbrace{P_I(A|x)}_{\text{A on 1st}} + \underbrace{(1 - P_I(A|x))P_I(A|x)}_{\text{B on 1st, A on 2nd}}$$

$$s_2(x) = P_I(A|x)$$

The marginal effect of x on grade share is:

$$s_1'(x) = 2(1 - P_I(A|x))P_I'(A|x)$$

$$s_2'(x) = P_I'(A|x)$$

Thus when $P_I(A|x) > \frac{1}{2}$, $s_1'(x) < s_2'(x)$, and the reverse when $P_I(A|x) < \frac{1}{2}$. Based on this example, the cleanest restaurants should get cleaner when re-inspections are removed, and the dirtiest to get dirtier.

Figure 2.12 shows the first-order condition (top half) and restaurant's chosen x states when removing re-inspections. As expected, the cleanest restaurants (those who chose $x \approx 6$ or below initially) improve their health states, while the dirtiest decline. As shown in the second column of Table 2.8, 311 food poisoning reports decline by 9.8-14.4% in this counterfactual. Improvements

come primarily from A and B restaurants, where food poisoning rates decline 9.6-14.5% and 8.3-10.4%, respectively.

False Negatives and False Positives One argument in favor of re-inspections is that they help avoid false positives (clean restaurants given a B grade or worse). Many restaurants complain about the variance in inspection scores from visit to visit, and re-inspections ensure that a restaurant must receive two substandard scores before being given a B or C.[42]

The counter to this argument, however, is that re-inspections generate a substantial number of false negatives, which are arguably more harmful from a public health perspective.

To shed light on this issue, I define an inspection cycle (initial + re-inspection) as a false positive if (i) the restaurant would be given an A if x were observed perfectly, and (ii) the restaurant is not given an A grade. False negatives are defined similarly. For example, for $x < \theta_{13}$, the probability of a false positive is:

$$f^+(x|\text{Re-Inspections}) = \left[1 - \Phi\left(\frac{\theta_{13} - x}{\sqrt{x}}\right)\right]^2$$
$$f^+(x|\text{No Re-Inspections}) = 1 - \Phi\left(\frac{\theta_{13} - x}{\sqrt{x}}\right)$$

Table 2.9 summarizes the rates of false positives and false negatives at baseline and upon removing re-inspections. As expected, there is a tradeoff when eliminating re-inspections: an increase in false positives versus a decline in false negatives. For restaurants with $x < \theta_{13}$, the average false positive rate increases from 7.2 percentage points to 10.6-14.9 percentage points.[43] However, the false negative rate declines dramatically upon eliminating re-inspections. Whereas restaurants with $x > \theta_{13}$ previously had a 60% probability of receiving an A grade, this rate drops to 20-35% when

[42] See, e.g., Willett-Wei 2014:

> A restaurant's grade is not wholly based on the quality of its food and the cleanliness of its kitchen...importantly, how strictly these specific rules are enforced (or noticed) on any given inspection varies by each individual health inspector.

[43] These numbers represent the average of each restaurant's probability of receiving a false positive. Due to different inspection frequencies, this is related to, but does not equal, the percent of inspections yielding false positives.

Table 2.9: Removing Re-Inspections: False Positive and Negative Rates

	With Re-Inspections	Without Re-Inspections
$\mathbb{E}(f^+(x))$	5.5	(9.6, 12.8)
$\mathbb{E}(f^+(x)\|x < \theta_{13})$	7.2	(10.6, 14.9)
$\mathbb{E}(f^-(x))$	13.6	(2.0, 5.0)
$\mathbb{E}(f^-(x)\|x > \theta_{13})$	60.2	(19.6, 35.4)

Description: False positive and negative rates at baseline (column 1) and upon removing re-inspections (column 2). Ranges represent minimum and maximum estimated outcomes for $\tilde{\rho} \in [0.1, 0.9]$.

eliminating re-inspections.

2.7.4 Removing Preferential Timing

In this counterfactual, I impose a uniform inspection cycle length. That is, instead of A-restaurants beginning their next inspection cycle after 12 months and C-restaurants after 4 months, every restaurant would begin a new inspection cycle X months after the conclusion of their current cycle. The exact value of X does not matter in this model, since the implied grade shares are invariant to a proportional increase in inspection cycle lengths (see Appendix B.2). For this counterfactual, inspector thresholds are set to baseline values and re-inspections are included.

A Simple Example

As in Section 2.7.3, removing preferential timing acts through the restaurant's first-order condition. Again it's easiest consider a simplified setting in which a restaurant can receive an A or B

grade. Their grade share function is then

$$\begin{aligned} s_A(x) &= \frac{t_A p(A|x)}{t_A p(A|x) + t_B p(B|x)} \\ &= \frac{1}{1 + \frac{t_B}{t_A}\frac{1-P(A|x)}{P(A|x)}} \\ &= \frac{1}{1 + \tau f(x)} \end{aligned}$$

where $\frac{t_B}{t_A} \equiv \tau$ is the relative length of a B-cycle vs. an A cycle, and $f(x) \equiv \frac{1-P(A|x)}{P(A|x)}$ is the odds ratio. At an interior optimum the first-order condition is:

$$\begin{aligned} \tilde{C}_r &= -s'_A(x) \\ &= \frac{\tau f'(x)}{(1 + \tau f(x))^2} \end{aligned}$$

Implicit differentiation establishes that $x^*(\tau)$ is decreasing in τ so long as

$$P(A|x^*(\tau)) > \frac{\tau}{1+\tau}$$

That is, for sufficiently low levels of τ at least, restaurants will respond to an increase in τ by investing more in their health practices. In New York, $P(A|x) > \frac{1}{2}$ for about 80% of restaurants, so health practices should, on average, improve after eliminating preferential timing (in this example, set $\tau = 1$).

The logic behind this result is apparent by taking an absurd example: suppose A grades lasted for 100 years and B grades for 1 day. So long as inspections aren't perfectly informative, restaurants can be relatively certainly that, even with poor health practices, they will quickly receive an A and reap the benefits of that grade for a long time. Increasing the relative length of the B cycle incentives restaurants to invest in order to continue to earn an A grade.

However, the effect of increasing α is non-monotonic. If A grades lasted for a day and B grades for 100 years, then restaurants would have little incentive to invest in receiving an A grade since

they will sooner or later be relegated, essentially forever, to a B. Thus finding the right balance between cycle times is an important design decision for the regulator.

Outcomes

Figure 2.13 shows the change in the restaurant's first-order condition when eliminating preferential timing. As the simple example above suggests, the marginal benefit of improving health practices improves for the vast majority of restaurants. Intuitively, removing preferential timing has the largest impact on restaurants that receive a range of grades with high probability. In New York, since so many restaurants consistently receive A's, this implies that the effects of eliminating preferential timing should be largest among the dirtiest restaurants, which is indeed what Figure 2.13 shows.[44]

As shown in the third column of Table 2.8, 311 food-poisoning reports decline by 9.9-13.3% in this scenario. Unsurprisingly, the improvements are particularly concentrated in B and C restaurants.

[44]A restaurant with a latent state of $x = 8$ (the bottom 3.7% of the distribution) has will spend roughly 69% of its time with an A grade at baseline, whereas when preferential timing is eliminated, an $x = 8$ restaurant would spend only 54% of its time with an A grade.

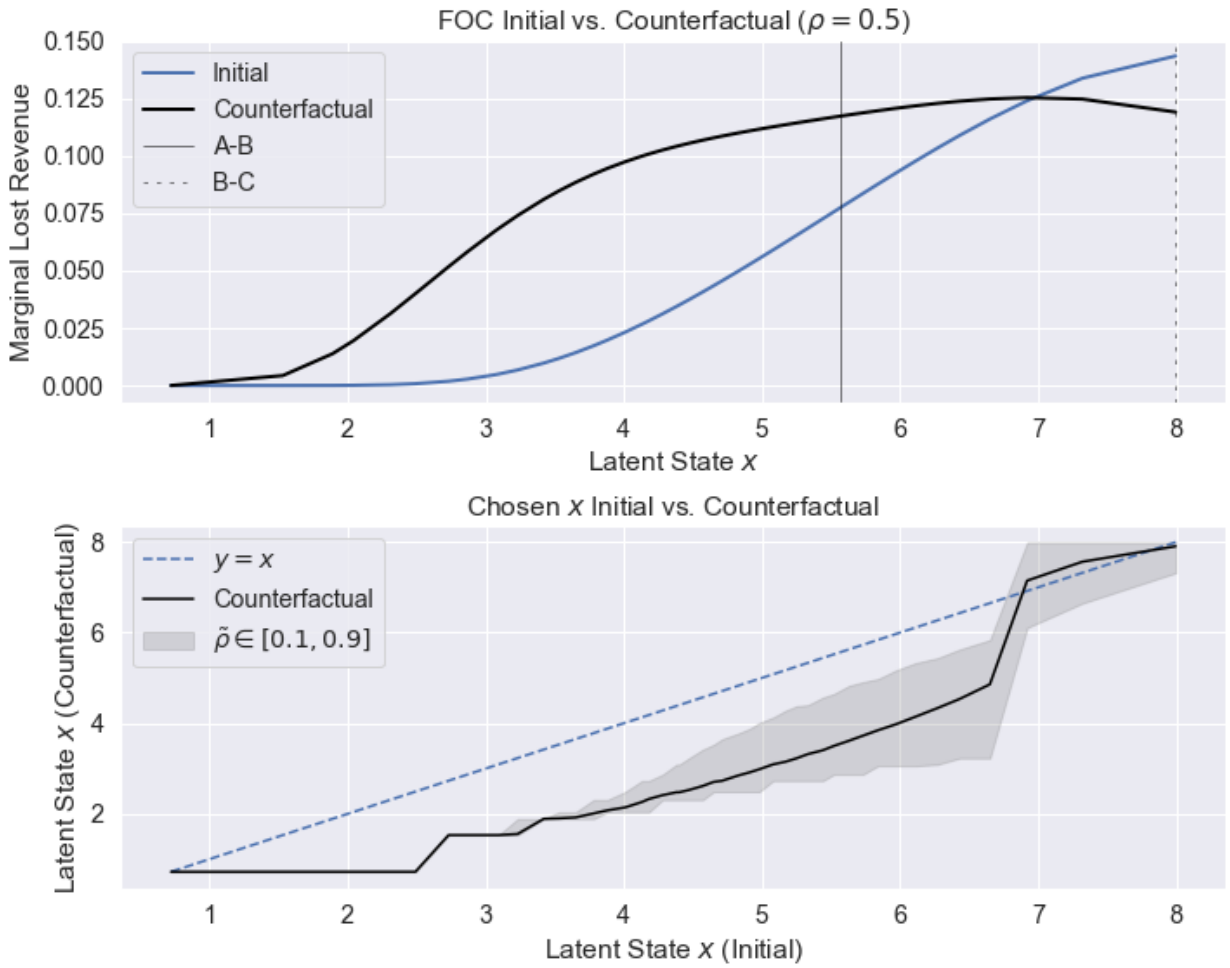


Figure 2.11: First-Order Condition and Health States when Eliminating Score Bunching.

Description: (Top chart) First-order condition $-s'_A(x) - s'_B(x)\tilde{\rho}$ initially (blue) and with $\theta_{13} = 3.5$ (black), shown for $\tilde{\rho} = 0.5$. A-B boundary represents the threshold at which an inspector reports an A versus a B. Since latent states are observed with normal error, a restaurant whose latent state is on the boundary will receive an A on 50% of inspections.
(Bottom chart) Initial and counterfactual latent states. Grey line represents the range of outcomes for $\tilde{\rho}$ in a grid of 15 points evenly spaced between 0.1 and 0.9. Black line represents the average of these outcomes.

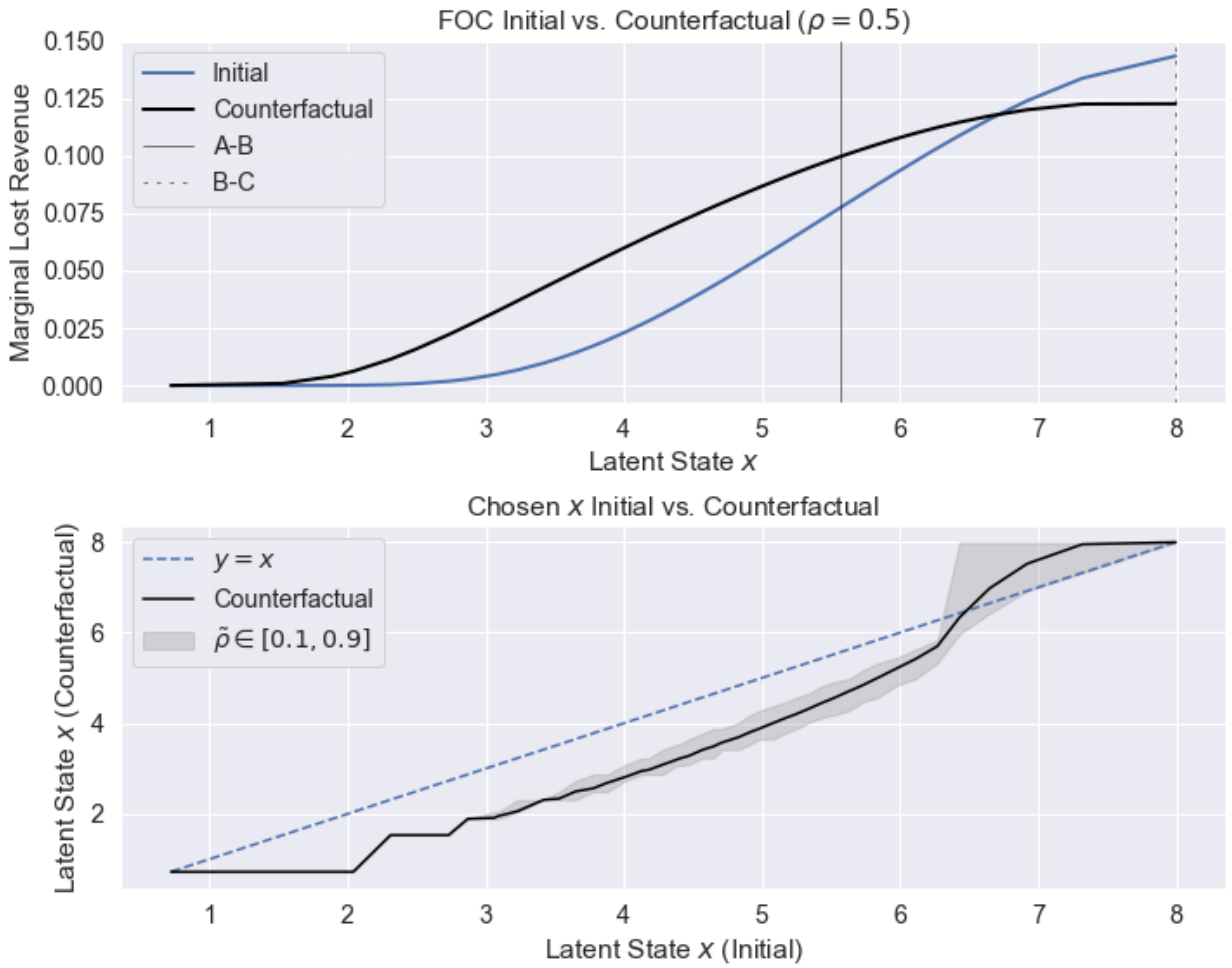


Figure 2.12: First-Order Condition and Health States when Eliminating Re-Inspections.

Description: (Top chart) First-order condition $-s'_A(x) - s'_B(x)\tilde{\rho}$ before (blue) and after (black) eliminating re-inspections, shown for $\tilde{\rho} = 0.5$. A-B boundary represents the threshold at which an inspector reports an A versus a B. Since latent states are observed with normal error, a restaurant whose latent state is on the boundary will receive an A on 50% of inspections.
(Bottom chart) Initial and counterfactual latent states. Grey line represents the range of outcomes for $\tilde{\rho}$ in a grid of 15 points evenly spaced between 0.1 and 0.9. Black line represents the average of these outcomes.

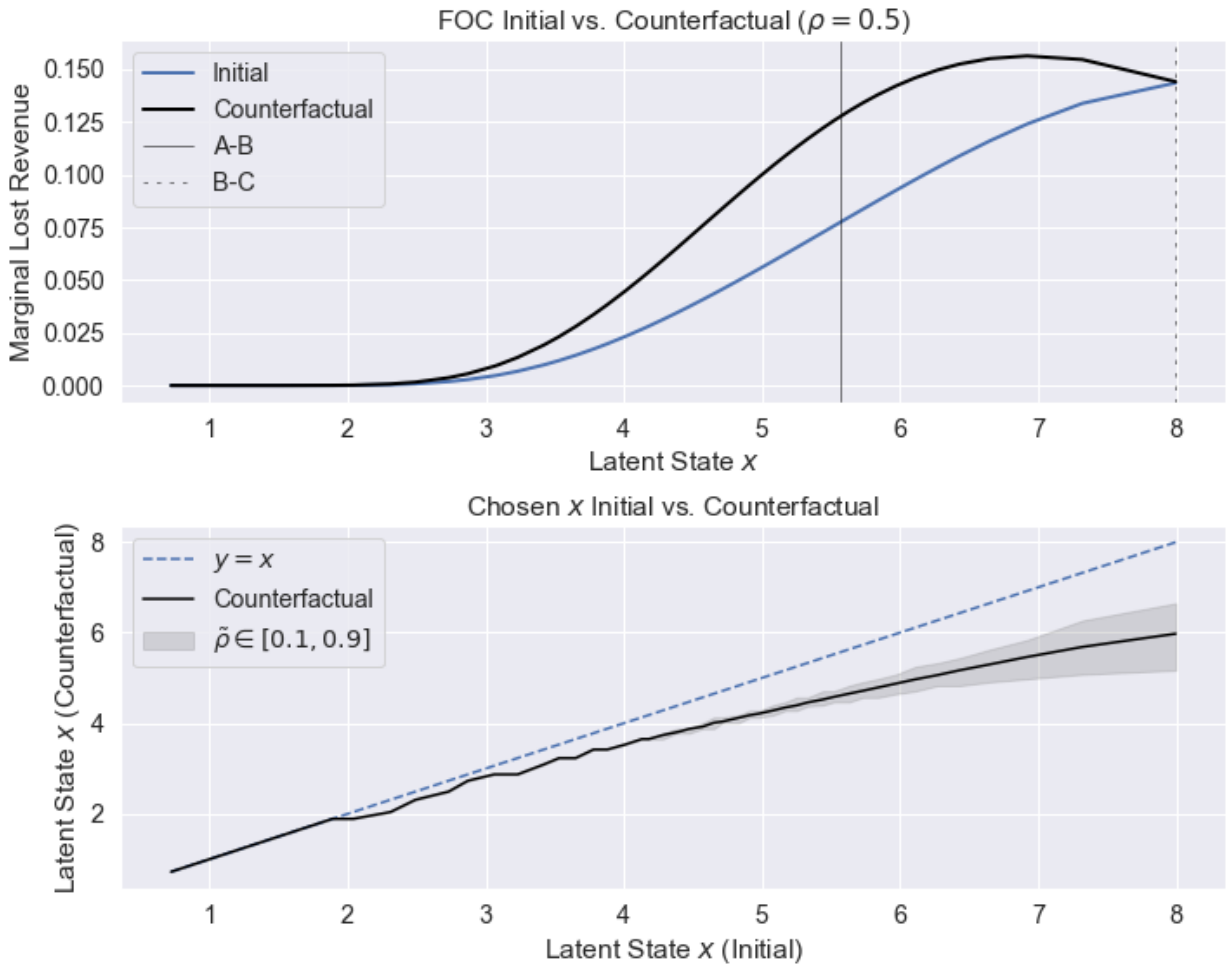


Figure 2.13: First-Order Condition and Health States when Eliminating Preferential Timing.

Description: (Top chart) First-order condition $-s'_A(x) - s'_B(x)\tilde{\rho}$ before (blue) and after (black) eliminating preferential timing, shown for $\tilde{\rho} = 0.5$. A-B boundary represents the threshold at which an inspector reports an A versus a B. Since latent states are observed with normal error, a restaurant whose latent state is on the boundary will receive an A on 50% of inspections.
(Bottom chart) Initial and counterfactual latent states. Grey line represents the range of outcomes for $\tilde{\rho}$ in a grid of 15 points evenly spaced between 0.1 and 0.9. Black line represents the average of these outcomes.

2.8 Conclusion

This chapter asks to what extent inspector preferences influence restaurant inspection scores in New York City. I argue health inspector behavior is consistent with a two-objective utility function: accurately reflecting restaurant food-poisoning risk, and avoiding giving restaurants B or C grades. Inspector desire to avoid B or C grades reduces restaurant incentives to invest in their health practices. Using a structural model of restaurant-inspector interaction, I find that if the A-B threshold were set by a social planner aiming to minimize a combination of food poisoning cases and restaurant compliance costs: (i) grading would be stricter than at present; and (ii) aggregate food poisoning cases would decline by 13-22%.

Moreover, I find two institutional features incentivize worse health practices for most restaurants: (i) re-inspections, in which restaurants that do not receive an A initially are granted a second opportunity to receive an A grade; and (ii) preferential inspection timing, in which restaurants that receive poor grades are inspected more frequently. Re-inspections give restaurants a second chance to receive an A grade, and reduce the marginal benefit of investing in health practices for most restaurants. Likewise, because bad grades are relatively short-lived, most restaurants' marginal benefit of investing in their health practices is lower under preferential timing than under a uniform timing policy. I find that removing re-inspections and preferential timing leads to a 10-14% and 10-13% decline in food poisoning cases, respectively.

Future research in this area would benefit from clarifying the source of inspector bias towards A grades. I find that inspector bias is largest in younger restaurants and those that have never been shut down by the Health Department. In addition, inspectors are more likely to cluster on the A side of the A-B boundary when they've had cases recently overturned through OATH hearings, suggesting that administrative burden and or the optics of having inspections overturned are important determinants of inspector behavior.

Chapter 3: Real-Time Inference: Evidence from MLB Umpires

3.1 Introduction

As discussed in the previous chapter, whether decision-makers appropriately weigh the information at hand when making judgements is a critical determinant of welfare. This applies not only to quality disclosure programs, but employment, lending, law-enforcement, and elsewhere. A substantial literature in psychology and behavioral economics has noted circumstances in which decision-makers may struggle to use information efficiently. Experts may place too little weight on relevant information, for example due to overconfidence (Barber and Odean 2001) or a desire to avoid contradictory evidence (Nickerson 1998). Alternatively, experts may put too much weight on irrelevant information (Tversky and Kahneman 1974).

In this chapter I use data on ball/strike calls from Major League Baseball umpires to investigate how decision-makers incorporate information into their judgements. Umpires are tasked with classifying pitches as balls (out of range of a batter's swing) or strikes (in range). Baseball offers an attractive waypoint between the lab and the "real world" to study questions of decision-making for several reasons. First, ball/strike calls are consequential decisions made by legitimate experts, but in a controlled setting that circumvents the omitted variable bias that plagues other studies of expert decision-making. Second the volume of data produced by Major League Baseball lets us track pitcher's tendencies as a function of the game state and construct the priors that a rational umpire would employ. Lastly, Major League Baseball's sophisticated pitch-tracking technology provides an accurate assessment of the quality of umpire decisions.

My primary finding is that umpires look remarkably Bayesian.[1] The logic of the result is simple. Umpires receive noisy measurements of pitch location: a typical pitch goes from the

[1]Several papers have found evidence of sports players exhibiting sophisticated behavior, such Chiappori et al. 2002's study of mixed-strategy equilibria in soccer penalty kicks.

pitcher's hand to the catcher's glove in roughly half a second, or two to three blinks of an eye.[2] When decision-makers receive noisy signals of the truth, there is room for prior information to inform posterior beliefs. If the umpire knows that a pitcher always throws strikes when he is behind in the count, his prior will push him towards calling a strike in those counts.[3] This is exactly what the data shows: for a fixed pitch location, umpires call more strikes when pitchers are behind in the count (and throwing many strikes) and vice versa when pitchers are ahead in the count (and throwing many balls). Of course umpires are not calculating Bayes' factors on the fly; rather, I expect they are pulling on their experience and leaning towards calling strikes in counts when they know pitchers will frequently throw one.

The tendency of umpires to call more strikes when pitchers are behind in the count and more balls when pitchers are ahead in the count has been observed several times before, and is typically presented as evidence of umpire bias, such as Green and Daniels 2014. Walsh 2010 and Mills 2014 cite this as evidence that umpires are compassionate, pushing outcomes in the favor a disadvantaged batter or pitcher. D. Chen et al. 2016 looks at it through the lens of the gambler's fallacy, with umpires less likely to call multiple strikes in a row. I instead show that this behavior can be given a rational explanation: umpires are Bayesian. Umpires are not compassionate; disadvantaged pitchers just throw a lot of strikes. Umpires don't suffer from a gambler's fallacy; after a strike, the next pitch tends to be further outside the strike zone, so umpires shrink their strike zone in response.[4]

I generate these results using data from Major League Baseball's Statcast system, an integrated camera and radar system that tracks the flight path of every pitch and batted ball. This allows me

[2]Along the way, the spin on the ball may cause the pitch to break up to a foot horizontally and vertically. The umpire needs to ascertain the location of the ball as it crosses home plate, all without having his vision obscured by the batter, the catcher, or the motion of the catcher's glove as he receives the ball. This is not an easy job.

[3]In an at-bat, if the pitcher throws four balls, the batter is granted first base, whereas after three strikes the batter is called out. A pitcher is "behind in the count" when there are more balls than strikes. The count of an at-bat is reported as "balls-strikes," so a 3-0 count means a count of three balls and no strikes.

[4]After writing the first draft of this chapter I became aware of an updated version of Green and Daniels 2014, Green and Daniels 2018, which comes to similar conclusions regarding umpire behavior. While similar, the analysis varies on several dimensions. My analysis allows for arbitrary count-specific preferences for false negatives vs. false positives and shows that, under the assumption of statistical discrimination, umpires exhibit little count-based variation in preferences. Green and Daniels 2018 assumes equal disutility from false negatives and false positives and shows that a model of statistical discrimination best describes umpire behavior. In addition, I highlight that the contours of the umpire's strike zone could potentially be used to identify a model in which both beliefs and count-specific preferences are allowed to vary, but that the results are inconclusive.

to measure the accuracy of umpire calls as a function of the game state. I couple this data with a recent model of expert decision-making developed in Chan et al. 2019 to infer umpire accuracy, their preferences for false positives versus false negatives, and the effect of count-specific pitch distributions (i.e. statistical discrimination) on umpire decisions.

At the heart of the model is an analysis of umpires' receiver operating characteristic ("ROC") curves, or their rate of true positives and false positives. When false positives are costly, umpires will locate at a point on their ROC with relatively few false positives, at the cost of more false negatives. Controlling for count-specific pitch distributions is critical in this analysis for two reasons. First, classification difficult varies by count. In a 3-0 count, the average strike is right down the middle (easy to classify) while the average ball is near the edge of the strike zone (hard to classify), naturally leading to more false positives. The second reason is statistical discrimination: the umpire's prior will lead him towards calling those marginal balls strikes, again leading to more false positives. Only if umpires call more strikes than expected based on these two factors would the model conclude they have a taste for calling strikes in 3-0 counts. The model suggests that variation in pitch location and priors explains the majority of variation in umpire decision-making, and that the count of the at-bat has little systematic relationship to an umpire's preferences.

This work is intimately related to the economics of discrimination. Economists typically distinguish two types of discrimination. Taste-based discrimination, as in Becker 1957, results when decision-makers make less favorable judgements for an otherwise equivalent group. Statistical discrimination, presented in Phelps 1972 and Arrow 1973, results when decision-makers accurately infer unobservable characteristics based on membership in a particular group. My primary finding is that umpire behavior is consistent with statistical discrimination.

Distinguishing taste-based from statistical discrimination is notoriously difficult but has important welfare and policy implications.[5] Becker 1957 argues that, so long as discrimination is not too rampant, taste-based discrimination can be competed away. Alternatively, while statistical discrimination is optimal from a signal-processing perspective, analyses like Arrow 1973 note that

[5]See Bertrand and Duflo 2016 for a recent review of literature on field experiments of discrimination.

it can generate self-fulfilling prophecies. This concern has prompted some governments to outlaw conditioning certain decisions on group status, for example the Equal Credit Opportunity Act or the Fair Housing Act in the United States.

The best field evidence on discrimination comes from settings with information on decision-maker judgements and detailed outcome data, the logic being that differences in judgements for observably identical groups implies discriminatory behavior. However, as discussed in Yinger 1996, if group status is correlated with important unobservables (such as social networks for repaying loans), even outcomes-based tests cannot separate statistical from taste-based discrimination.[6] The benefit of the MLB setting is that I receive exact measures of the variables umpires are supposed to condition their decision on, reducing the fear that omitted variable bias is contaminating the results.

The model assumes umpires act like statistical decision-makers and shows that, under this assumption, they exhibit little taste-based discrimination. While I find this interpretation cleaner than that of an umpire whose preferences vary dramatically from count-to-count (or an umpire whose behavior is some combination of these extremes), identifying the extent to which umpires statistically discriminate is difficult. As discussed in Bohren et al. 2019's recent working paper, most empirical work can only identify taste-based discrimination under the assumption that decision-makers statistically discriminate with accurate beliefs. If beliefs are inaccurate, there are a continuum of discriminatory preferences that can rationalize observed data. In the Appendix I discuss how the shape of an umpire's strike zone can theoretically be used to identify the extent of statistical discrimination in the model. The basic idea is simple: if umpire's are statistical discriminators, then asymmetric pitch distributions will tend to produce asymmetric strike zones, whereas taste-based discrimination (variation in preferences for false positives versus false negatives) produces radially symmetric changes in an umpire's strike zone. Unfortunately this test is low-powered, and I am not able to resolve whether umpires are in fact statistical discriminators.

The chapter proceeds at follows. Section 3.2 outlines the basic logic of Chan et al. 2019's

[6]Examples of work that uses outcome-based data to study questions of discrimination include Knowles et al. 2001 (police vehicle searches) and Pope and Sydnor 2011 (peer-to-peer lending).

decision model. Section 3.3 provides further background on Major League umpires and the Statcast system. Section 3.4 provides evidence that umpire decision-making varies substantially from count-to-count. Section 3.5 develops and estimates the empirical model and shows that most of the variation in umpire behavior is due to count-specific pitch location and statistical discrimination, not variation in preferences. Section 3.6 concludes.

3.2 A Model of Expert Decision-Making

To unpack the relative importance of statistical discrimination versus umpire preferences, I use a model of expert decision-making outlined in Chan et al. 2019. This section briefly discusses the underpinnings of this model.

Consider a decision-maker ("DM") tasked with making a binary classification. DMs are statistical decision-makers: they receive an informative signal about the state of the world and then decide whether to report positive or negative based on their signal.

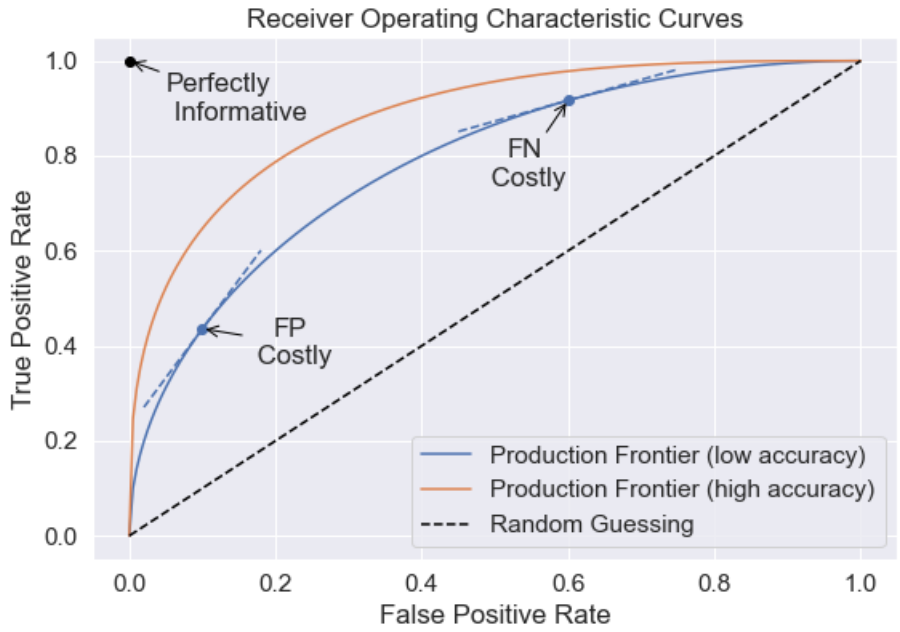
A standard tool for evaluating the DM's effectiveness is a *receiver operating characteristic curve*, or ROC curve. An ROC curve plots the DM's true positive rate against their false positive rate. As shown in Figure 3.1, the point (0,1) corresponds to a perfect classifier, whereas a point (p, p) corresponds to the performance of a classifier that simply guesses "positive" with probability p. ROC curves admit a partial order in the spirit of Blackwell 1953: one ROC curve is more accurate than another if it sits strictly to the northwest of another.

Economically, ROC curves correspond to production frontiers, or the possible error profiles a decision-maker can produce given a particular decision-making technology. Where a DM locates along an ROC curve depends on a tangency condition between their indifference curve and the ROC curve. If the DM's utility function is $u(FP, FN)$, their first-order condition is

$$\frac{N}{P}\frac{u_{FP}}{u_{FN}} = \frac{dTPR}{dFPR}$$

where N and P are the population number of negatives and positives, respectively. That is, after

Figure 3.1: Sample ROC Curves



weighting for population frequency, the optimal policy for DMs balances the relative costs of false positives and false negatives against the marginal increase in true positives along the ROC.

Under modest assumptions about the DM's decision-making technology, the true positive rate is a concave function of the false positive rate.[7] In this case, the first-order condition above implies the following comparative statics:

- As N increases or P decreases, true and false positive rates decrease; i.e. when the negative state is more common DM's report more negatives
- As u_{FN} decreases or u_{FP} increases, true and false positives decrease; i.e. when false negatives are less costly DMs report more negatives

ROC curves provide a substantial amount of information about a DM's accuracy and preferences for false positives versus false negatives. In Section 3.5 I discuss how to extend this frame-

[7] In particular, the DM need only be able to randomize between feasible points. Suppose $a = (FPR_1, TPR_1)$ and $b = (FPR_2, TPR_2)$ are two points on the DM's ROC curve. If the DM can randomly pick between the decision rule that generates a and b, all convex combinations of a and b are feasible outcomes, and so must sit weakly below the ROC curve.

work to include decision-specific observables, capturing both taste-based and statistical aspects of variation in DM behavior across settings.

3.3 Empirical Setting: MLB Umpire Decisions

The empirical work uses data on ball/strike calls by Major League Baseball umpires. While this chapter is not focused on the rules or strategy of baseball per se, a bit of context is useful to think through the umpire's decision-making process.

The key interaction in baseball is between a batter and a pitcher. Batters attempt to reach base by hitting the ball into the field of play, while pitchers attempt to prevent this. Each at-bat is overseen by a home plate umpire, whose task (among others) is to categorize each pitch as a "ball" (outside the strike zone) or a "strike" (inside the strike zone).[8] If an at-bat reaches four balls, the batter is granted first base, and if an at-bat reaches three strikes, the batter is declared out. In 2019, Major League Baseball defined the strike zone as follows:

> The STRIKE ZONE is that area over home plate the upper limit of which is a horizontal line at the midpoint between the top of the shoulders and the top of the uniform pants, and the lower level is a line at the hollow beneath the kneecap. The Strike Zone shall be determined from the batter's stance as the batter is prepared to swing at a pitched ball.[9]

Major League Baseball is an attractive setting to study models of expert decision making for two reasons. First, umpires face a clear binary classification problem: whether to call a pitch a strike or not. Second, for more than a decade now, Major League Baseball has captured exceptionally precise information on every pitch thrown. This data not only provides a high-powered way to investigate how game conditions influence umpire decision-making, but it also an independent measure of whether a pitch was a strike or not. This lets me easily classify decisions as false

[8] Pitches the batter makes contact with, such as foul balls, are automatically ruled strikes.
[9] 2019 Official Baseball Rules, p. 153; see MLB Playing Rules Committee 2019.

positives or false negatives, avoiding the econometric difficulties created by standard one-sided selection models where the true outcome is only revealed for a subset of cases.[10]

PITCHf/x and Statcast Data Since 2008, Major League Baseball has used tracking software to generate information about almost every pitch thrown. The first generation of this technology, used from 2008 to 2014, was known as PITCHf/x. The PITCHf/x system used two cameras to sample the ball's flight path about twenty times between when it left a pitcher's hand and when it arrived at home plate. Based on these samples, the complete path of the ball could be reconstructed under a constant acceleration assumption, yielding detailed pitch-specific data: its speed, movement in flight, spin rate, and — most importantly for this chapter — its location upon crossing home plate.[11]

Major League Baseball's second generation pitch-tracking software, Statcast, was adopted in 2015. Statcast is a more holistic data-gathering technology than PITCHf/x, using an integrated camera and radar system to track batted balls and player movements in addition to pitch data.[12] Presently, however, I am primarily interested in Statcast's pitch location data, which is directly comparable to PITCHf/x's. While there have been reports of inconsistencies between PITCHf/x and Statcast data (e.g., Schifman 2018), the primary results hold in both the PITCHf/x and Statcast eras, and exhibit no noticeable discontinuities at the time of the technology switch.

PITCHf/x and Statcast data are publicly available through Baseball Savant, Major League Baseball's clearinghouse for Statcast data.[13] In addition to pitch-specific data, PITCHf/x and Statcast collect general game information (data, home and away teams) as well as pitch-specific features (score, count, pitcher, batter, runners on base, etc.), allowing me to recreate the circumstances of an umpire's decision problem at a granular level.

[10]For example, in Chan et al. 2019's study of pneumonia diagnoses, they observe false negatives (a patient with a negative diagnosis who returns in a few days and receives a positive diagnosis) but do not directly observe false positives.

[11]For more information on the PITCHf/x system, see Fast 2010.

[12]See `http://m.mlb.com/glossary/statcast`.

[13]See `https://baseballsavant.mlb.com/statcast_search`

Defining a Strike Statcast reports a pitch's location when crossing the plate along both the horizontal x axis and the vertical z axis. Both measures are reported in feet. The x axis is centered at 0 (the middle of home plate), while the 0 of the z axis corresponds to the ground.

To determine whether a pitch is a strike, I first check if any portion of the ball passes over home plate. Home plate is 17 inches wide, and a baseball is a sphere approximately 2.90 inches in diameter.[14] Therefore a pitch passes over home plate if

$$|x_i| < \frac{17 + 2.90}{2}$$

Next, I determine whether the pitch's vertical location sits between "the midpoint between the top of the shoulders and the top of the uniform pants" and "the hollow beneath the kneecap." Statcast camera operators are trained to adjust the vertical limits of the strike zone by batter, and report these limits in the "sz_top" and "sz_bot" fields. These fields are noisy, though. Figure 3.2 shows the distribution of strike zone heights for Aaron Judge (at 6 foot 7 inches, one of the tallest players in Major League Baseball), and Jose Altuve (at 5 foot 6 inches, of the shortest) in their 2019 at-bats. While Altuve's strike zone is accurately lower than Judge's, the limits vary up to half a foot for the same batter. To correct for this noise, I create batter-specific strike zones by taking the average values of "sz_top" and "sz_bot" for each batter.

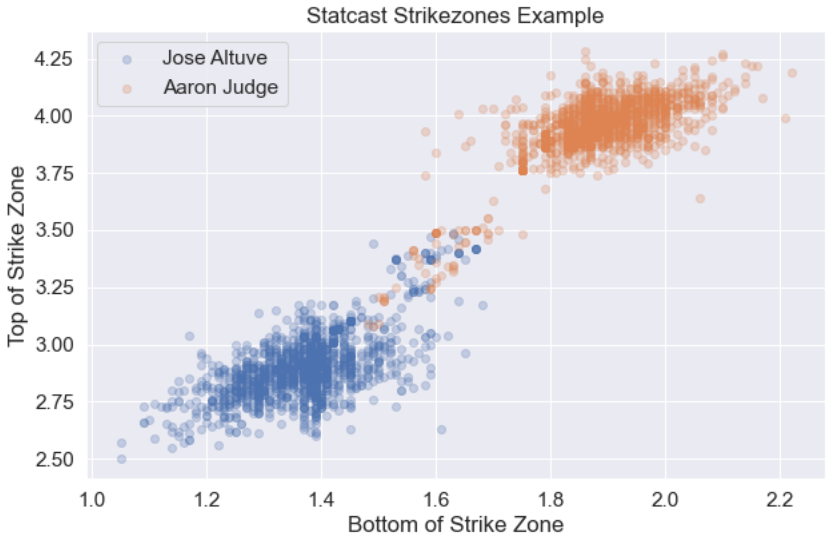
3.4 Model-Free Evidence: Umpire Sensitivity to Count

One of the most striking features of umpire decision-making is how their behavior varies depending on the count of an at-bat. Figure 3.3 plots aggregate true positive and false positive rates by count for the 2019 season. As a reminder, a false positive is a pitch that is called a strike but is actually a ball, while a true positive is a strike that is called a strike.

The first thing to notice in Figure 3.3 is the magnitude of the heterogeneity: false positive

[14] 2019 Official Baseball Rules, Sections 2.02 ("Home Base") and 3.01 ("The Ball"); see MLB Playing Rules Committee 2019

Figure 3.2: Statcast Vertical Strike Zone Limits



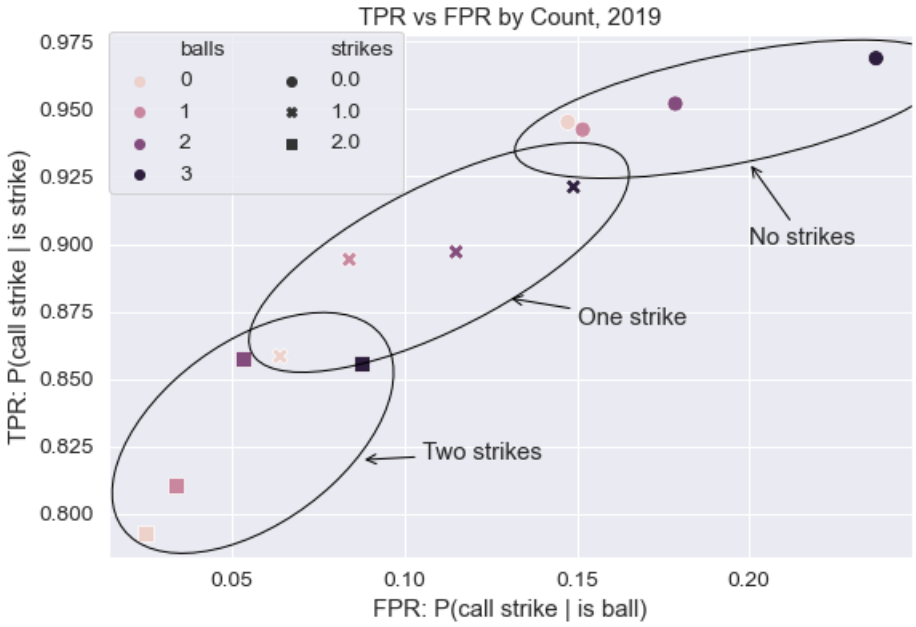
Description: Strike zone limits (in feet) using Statcast's "sz_bot" and "sz_top" fields for Aaron Judge and Jose Altuve. Each dot represents a pitch in the 2019 MLB season.

and true positive rates vary by more than 15 percentage points across counts.[15] Closer inspection reveals a nearly linear relationship between count and false positive rates: false positive rates increase with the number of balls and decrease with the number of strikes. This leads to strikingly different behavior at the extremes of an 0-2 versus a 3-0 count. Put simply, umpires call far fewer strikes in 0-2 counts and far more strikes in 3-0 counts.

Other studies, such as Green and Daniels 2014, have observed these count effects and interpret this behavior as evidence of umpire bias. The framework in Section 3.2 provides a language to discuss these effects. If umpires find false negatives (a mistaken ball call) more costly in no-strike counts, they would locate further northeast on their ROC curve in no-strike counts. That is, the umpire behavior noted in Figure 3.3 could be explained as arising from count-specific variation in the false negative utility parameter β. If this variation arises due to factors the social planner does not want the umpire to consider (such as social pressure, or a desire not to make decisive calls), umpire behavior may be socially harmful.

[15]The average standard deviation of the true and false positive rates are 0.52 and 0.28 percentage points, respectively.

Figure 3.3: Umpire True and False Positive Rates by Count



Description: Aggregate true positive and false positive rates by count for all pitches not swung at in the 2019 MLB season.

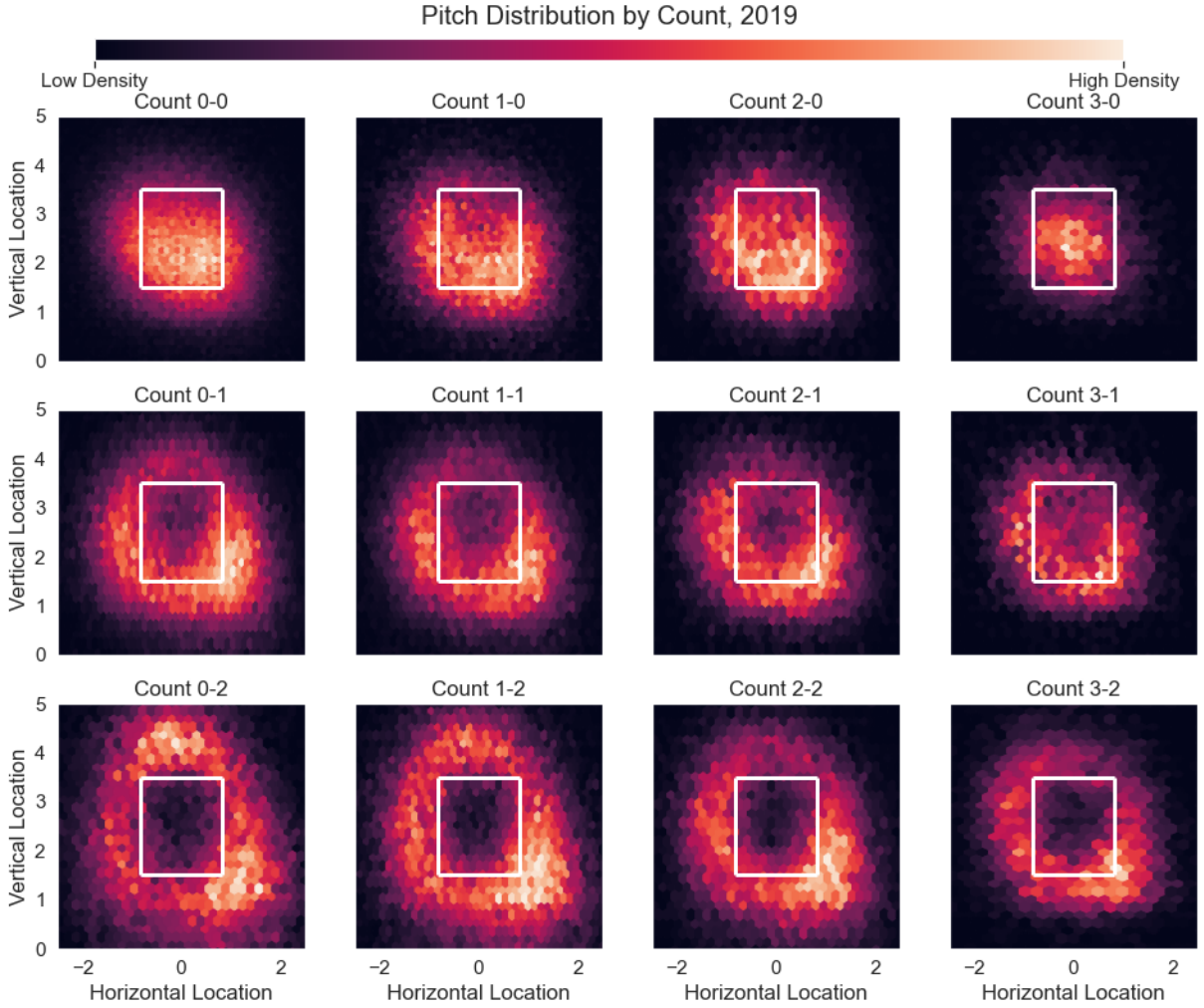
However, the model in Section 3.2 thinks of umpires as statistical decision makers. Since the current count is observable to umpires, the umpire's optimal decision rule will depend on a pitch's *observable risk*, or the count-specific distribution of pitch locations. Observable risk turns out to vary substantially across counts. Consider Figure 3.4, which shows the distribution of pitch locations by count in 2019. The top-right corner shows that, in 3-0 counts, pitchers typically throw the ball right down the middle of the plate. For a marginal pitch in a 3-0 count, the count-specific prior will push the umpire's posterior pitch location towards the center of the strike zone, making him more like to call the pitch a strike.

The opposite effect prevails in 0-2 counts: here, pitchers throw pitches outside the strike zone in the hopes that a batter will chase a bad pitch. On a marginal pitch in an 0-2 count, the count-specific prior will push the umpire's posterior pitch location away from the center of the strike zone, making him less likely to call the pitch a strike. These effects are exactly in line with Figure 3.3: umpires call fewer strikes in high-strike counts and more strikes in high-ball counts. This gives

a bias-free explanation of umpire behavior: umpires recognize that their signals of pitch location are imperfect, and err on the side of calling strikes in counts when pitchers almost always throw strikes. To fully disentangle bias from observable risk requires a model – which I turn to next — but the preceding analysis highlights the important role observable risk plays in determining the output of such a model.[16]

[16]There is also variation in how easy it is to classify a typical ball or strike across counts. For example, in a 3-0 count, the typical strike is in the middle of the strike zone and quite easy to classify. Since analyses like Green and Daniels 2014 find sizable count effects even after controlling for pitch location, I have focused attention in the preceding discussion on count-specific pitch distributions.

Figure 3.4: Pitch Location by Count



Description: Hexagonal density plot of the 2019 pitch distribution by count, where the color of each hexagon correspond to that location's count-specific frequency. The sample includes all pitches that umpires are required to make a ball/strike call on (i.e. it excludes batted balls). The number of divisions of the x- and z-axes is proportional to the cube root of the sample size. The white overlay represents a strike zone with vertical limits of 1.5 and 3.5 feet.

3.5 A Model of Umpire Decision Making

This section extends the model of Chan et al. 2019 to a binary classification problem with a two-dimensional signal and case-specific observable risk. I then estimate this model using data on MLB umpire decisions from 2008 to 2019.

Structure of the Model Pitch i is thrown at coordinates (x, z) (I suppress further i subscripts for legibility). The umpire observes a noisy signal of the pitch's location, $(\hat{x}, \hat{z})$, where:

$$\hat{x} = x + \epsilon_x$$

$$\hat{z} = z + \epsilon_z$$

Associated with the pitch is an information set $\mathcal{I}$. This information set could be extensive: the current count, score, and inning; the teams involved; attendance; the current pitcher, etc. In estimation I take $\mathcal{I}$ to be the current count.

For computational tractability I assume the disturbances (ϵ_x, ϵ_z) satisfy:

$$\begin{pmatrix} \epsilon_x \\ \epsilon_z \end{pmatrix} = \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_x^2(\mathcal{I}) & 0 \\ 0 & \sigma_z^2(\mathcal{I}) \end{pmatrix}\right),$$

where the parameters σ_x, σ_z are allowed to vary depending on the information set $\mathcal{I}$.

Given an observed location $(\hat{x}, \hat{z})$, the posterior location distribution is:

$$\begin{aligned} P(x, z|\hat{x}, \hat{z}, \mathcal{I}) &\propto P(\hat{x}, \hat{z}|x, z)P(x, z|\mathcal{I}), \\ &\propto \varphi\left(\frac{\hat{x} - x}{\sigma_x(\mathcal{I})}\right)\varphi\left(\frac{\hat{z} - z}{\sigma_z(\mathcal{I})}\right)P(x, z|\mathcal{I}) \end{aligned}$$

where $P(x, z|\mathcal{I})$ is the observable risk, i.e. the prior distribution of pitch locations given information set $\mathcal{I}$.

The posterior location distribution can be used to calculate the probability a pitch is truly a

strike for a given observed location:

$$P(s|\hat{x},\hat{z},\mathcal{I}) = \int_{x,z} \mathbf{1}(s|x,z)P(x,z|\hat{x},\hat{z},\mathcal{I})$$

Given an observed location $(\hat{x},\hat{z})$, the umpire's problem is to determine whether to call a ball or a strike. Suppose the umpire's utility in each case is:

$$u(s|\hat{x},\hat{z},\mathcal{I}) = -\underbrace{(1 - P(s|\hat{x},\hat{z},\mathcal{I}))}_{FP(\hat{x},\hat{z},\mathcal{I})}$$
$$u(b|\hat{x},\hat{z},\mathcal{I}) = -\beta(\mathcal{I})\underbrace{P(s|\hat{x},\hat{z},\mathcal{I})}_{FN(\hat{x},\hat{z},\mathcal{I})}$$

Simple algebra shows that the umpire will call a strike so long as:

$$P(s|\hat{x},\hat{z},\mathcal{I}) \geq \frac{1}{1+\beta(\mathcal{I})}$$

The more costly false negatives are (calling ball when the pitch is a strike), the more often the umpire will call strikes.

Define $P^u(s|x,z,\mathcal{I})$ as the probability an umpire calls a strike given a true location (x,z). Integrating over the range of observed signals and applying Bayes' Rule gives:

$$P^u(s|x,z,\mathcal{I}) = \int_{\hat{x},\hat{z}} P(\hat{x},\hat{z}|x,z,\mathcal{I})\mathbf{1}\left\{P(s|\hat{x},\hat{z},\mathcal{I}) \geq \frac{1}{1+\beta(\mathcal{I})}\right\} \tag{3.1}$$

For empirical work I use the Statcast measurements for (x,z). This is not exactly correct, as Statcast may have measurement error of its own; indeed, Major League Baseball's testing suggests Statcast has a mean absolute error of approximately one half inch, although in isolated circumstances the error may be larger.[17] There is no theoretical problem with incorporating Statcast measurement error into the model: for a Statcast measurement $(x_s,z_s) = (x+\epsilon_s, z+\epsilon_s)$, the likelihood

[17] See Fast 2011.

equation becomes:

$$P^u(s|x_s, z_s) = \int_{x,z} P(x, z|x_s, z_s) P^u(s|x, z)$$

However, the additional integral increases the computational burden substantially. Moreover, since the model suggests Statcast is considerably more accurate umpires — it estimates umpire signals had a mean absolute error of approximately 3 inches in 2019 — I do not expect that unmodeled measurement error is a significant source of estimation bias.[18]

Estimation I use a Bayesian estimation approach based on Equation 3.1. For a given pitch i and a vector of parameters $\theta = (\sigma_x, \sigma_z, \beta)$, the probability of observing a strike call $s_i \in \{0, 1\}$ is:

$$P_i^u(\theta) = P^u(s_i|x_i, z_i, \mathcal{I}_i, \theta)^{s_i} (1 - P^u(s_i|x_i, z_i, \mathcal{I}_i, \theta))^{1-s_i} \tag{3.2}$$

Let $\mathcal{D}_\mathcal{I}$ denote the location and strike data for all pitches that share information set $\mathcal{I}$. Given this data, the posterior distribution of θ is given by:

$$\begin{aligned} P(\theta|\mathcal{D}_\mathcal{I}) &\propto P(\mathcal{D}_\mathcal{I}|\theta) P(\theta) \\ &= \left(\prod_{i \in \mathcal{D}_\mathcal{I}} P_i^u(\theta) \right) P(\theta) \end{aligned}$$

Thus, for a given prior $P(\theta)$, I can estimate the information-set-specific posterior distribution of θ as follows:

1. For each year of Statcast (or PITCHf/x) data, split pitches into 12 information sets based on count.

2. For each count, calculate $P(\theta|\mathcal{D}_\mathcal{I})$ on a large grid of feasible values for $(\sigma_x, \sigma_z, \beta)$.[19] I assume a uniform prior over these grid points.

[18] In 2019, the model suggests $(\sigma_x, \sigma_z) \approx (0.25, 0.15)$, measured in feet. The mean of the absolute value of a normal distribution with covariance matrix $\Sigma = \begin{bmatrix} 0.25^2 & 0 \\ 0 & 0.15^2 \end{bmatrix}$ is approximately 0.25 feet, or 3 inches.

[19] We use a 2,880 point ($12 \times 12 \times 20$) grid of values for $(\sigma_x, \sigma_y, \beta)$, with $\sigma_x \in [0.2, 0.6]$, $\sigma_z \in [0.05, 0.4]$ and $\beta \in [0.5, 3]$.

3. Sample from the posterior distribution using the Metropolis-Hastings algorithm, using the grid parameters that maximize the posterior likelihood as the starting point.[20]

4. Construct posterior means and confidence intervals for $(\sigma_x, \sigma_z, \beta)$ from the Metropolis Hastings samples.

5. To highlight the importance of controlling for observable risk, I rerun the analysis above, except without splitting the data based on count.

While computationally expensive, a Bayesian approach is particularly convenient for this model due to the non-differentiability of the likelihood function with respect to β, which complicates standard maximum likelihood techniques.

Results Figure 3.5 shows the results of the estimation procedure. Each dot represents the posterior mean of a count-specific parameter in a given year.[21] I find that umpires' signals are more precise along the z-axis than the x-axis, although precision along the x-axis has nearly doubled since 2008.[22] In 2019, the the typical umpire error was 1-2 inches along the z-axis and 2-3 inches along the x-axis.[23]

Estimates of β suggest umpires place approximately 20-40% more weight on false negatives than false positives — umpires find mistaken ball calls more costly — although there is variation in $\hat{\beta}$ across counts. In order to understand the impact of count on umpire preferences, Table 3.1 projects the posterior means onto count data and year fixed effects using OLS. As shown in the last

[20] I take 5000 draws for each count-year. The transition rule is a normal random walk with standard deviation 0.01 in all dimensions.

[21] The average posterior standard deviation for $(\sigma_x, \sigma_z, \beta)$ are $(0.004, 0.003, 0.013)$.

[22] Several media outlets have noted this improvement in umpire performance, which many attribute to the feedback from Statcast and PITCHf/x. Consider this quote from Dusty Dellinger, a former MLB umpire cited in Davis and Lopez 2015:

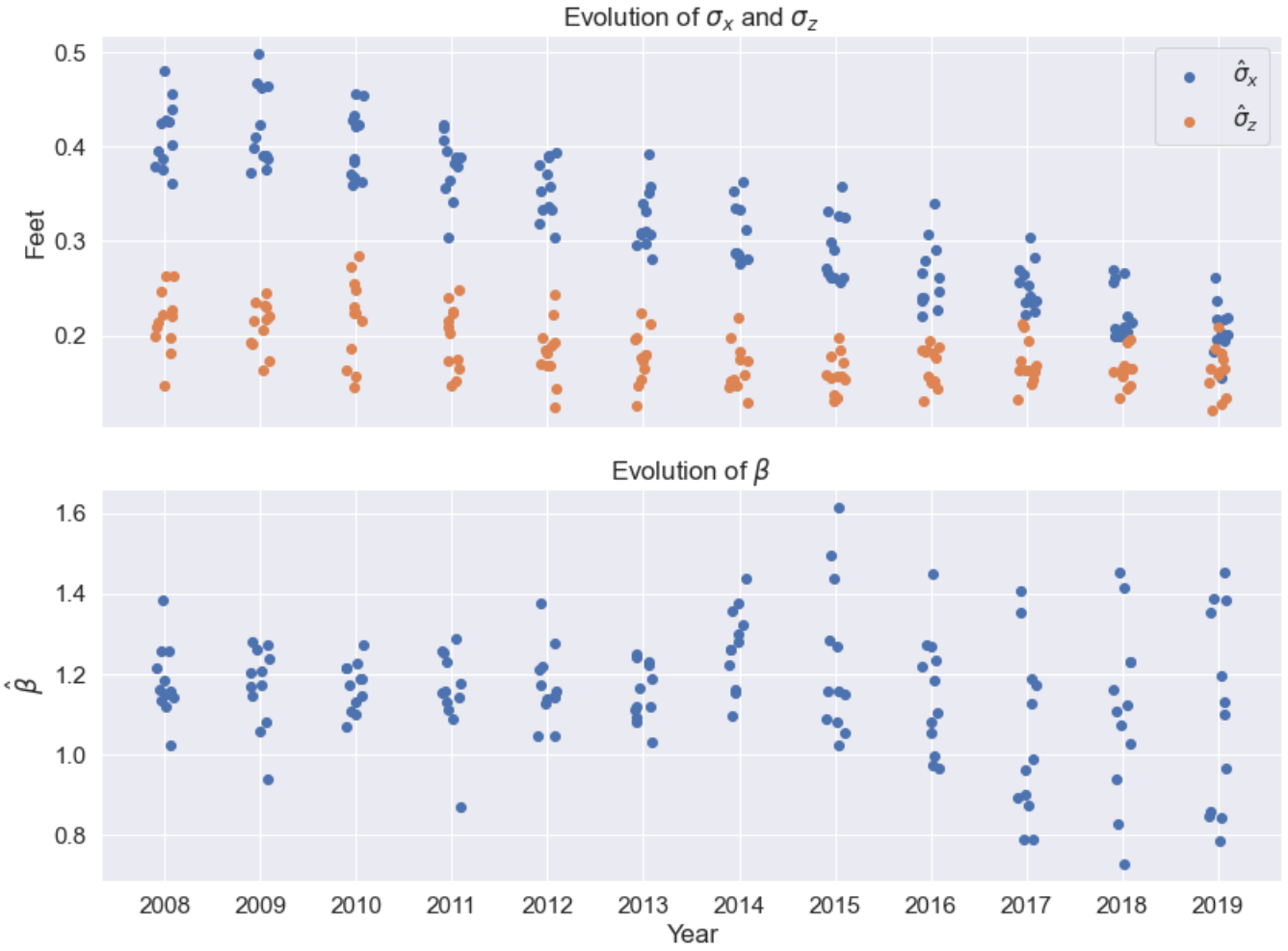
> It was amazing how my perspective of the strike zone changed when I got this technology. I thought pitches were on the plate, until you get that data back. You see that some of those pitches were not on the plate. It wasn't something that was done intentionally. It was just your perception of the strike zone. I was able to quickly make adjustments based on having that information, which was huge to me.

[23] The fact that umpires err more along the horizontal axis is not too surprising. To avoid having their vision obscured by the catcher, most umpires center themselves along the inside corner of the plate. This leads to a noticeable asymmetry in the data: umpires tend to be more accurate on inside pitches, and less accurate on outside pitches.

two columns, balls and strikes do appear to influence umpire's utility weights, but the fit is noisy. I estimate that incrementing the count by one ball increases the relative cost of a false negative by 0.6 percentage points. This is in line with umpires preferring to call strikes in high-ball counts.

The coefficient on strikes suggests that relative cost of a false negative falls by 5.7 percentage points for each strike, consistent with umpires preferring to call balls in high-strike counts. However, it is worth keeping in mind that the standard errors in Table 3.1 only reflect sampling error, not the uncertainty inherent in the estimated parameters. Since a more careful quantification of uncertainty would stretch the confidence interval even further, the clearer conclusion is that I do not find strong evidence that umpires dramatically change their preferences in response to count.

Figure 3.5: Estimated Umpire Parameters



Description: Note: Estimated parameters $(\sigma_x, \sigma_z, \beta)$ by year. Each dot represents the posterior mean of a parameter in a different count. There are 12 dots per parameter-year. To distinguish points more easily, x-values have been given a random jitter, uniform between -0.1 and 0.1.

Table 3.1: Projecting Umpire Parameters onto Count Data

	σ_x		σ_z		β	
Intercept	0.309^{***}	0.402^{***}	0.177^{***}	0.210^{***}	1.209^{***}	1.232^{***}
	(0.013)	(0.010)	(0.005)	(0.007)	(0.025)	(0.046)
Balls	0.001	0.001	-0.010^{***}	-0.010^{***}	0.006	0.006
	(0.006)	(0.002)	(0.002)	(0.001)	(0.011)	(0.011)
Strikes	0.009	0.009^{***}	0.020^{***}	0.020^{***}	-0.057^{***}	-0.057^{***}
	(0.008)	(0.003)	(0.003)	(0.002)	(0.015)	(0.014)
Year FE	N	Y	N	Y	N	Y
Adj R^2	-0.004	0.830	0.316	0.668	0.083	0.156
N			144			

$^{***}p < 0.01,^{**}p < 0.05,^{*}p < 0.10$

Description: OLS regressions of estimated coefficients $(\sigma_x, \sigma_z, \beta)$ on count variables and year fixed-effects.

Figure 3.6 demonstrates another method for decomposing the observed differences in true and false positive rates across counts. In the model there are three factors that can influence true and false positive rates across counts:

1. Variation in umpire parameters
2. Variation in pitch location
3. Variation in the umpire's prior

The second and third factors are related, but not the same. Variation in pitch location refers to the fact that pitches are easier to classify in certain counts. In 3-0 counts, for example, the typical strike is in the middle of the plate and is easy to classify, leading to fewer false negatives in these counts. Variation in the umpire's prior refers to the fact that, for a fixed location, the umpire will be more likely to call a pitch a strike in some counts versus others. Using 3-0 counts again, the fact that most pitches are in the middle of the strike zone causes the umpire to expand his strike zone

relative to other counts.

Figure 3.6 performs such a decomposition. To create this figure, I estimate four versions of the model. In the first model, I shut down variation from all three of the sources noted above. That is, I estimate what an umpire's true and false positive rates would be if their parameters were set to the mean of the estimated 2019 parameters and the pitch distribution in each count equalled the 2019 aggregate pitch distribution. In the second model I allow umpire preferences to vary by count. In the third model I allow the pitch distribution to vary by count, but assume umpires use the aggregate pitch distribution as their prior when evaluating pitches. The final model corresponds to the full model fitted to 2019 data, where umpire priors reflect the count-specific pitch distribution.

The path of a particular count, traces out the effect of parameters, pitch location, and priors on umpire performance. Consider the false positive rate for 3-0 counts. I estimate that umpires' preferences actually pushed them towards calling slightly fewer false positives in 2019, but that the pitch distribution on 3-0 counts more than overcame that effect. Pitch distribution is also the dominant effect for false positives in 3-0 counts.

One simple way to estimate the relative importance of these three factors is using the decomposition:

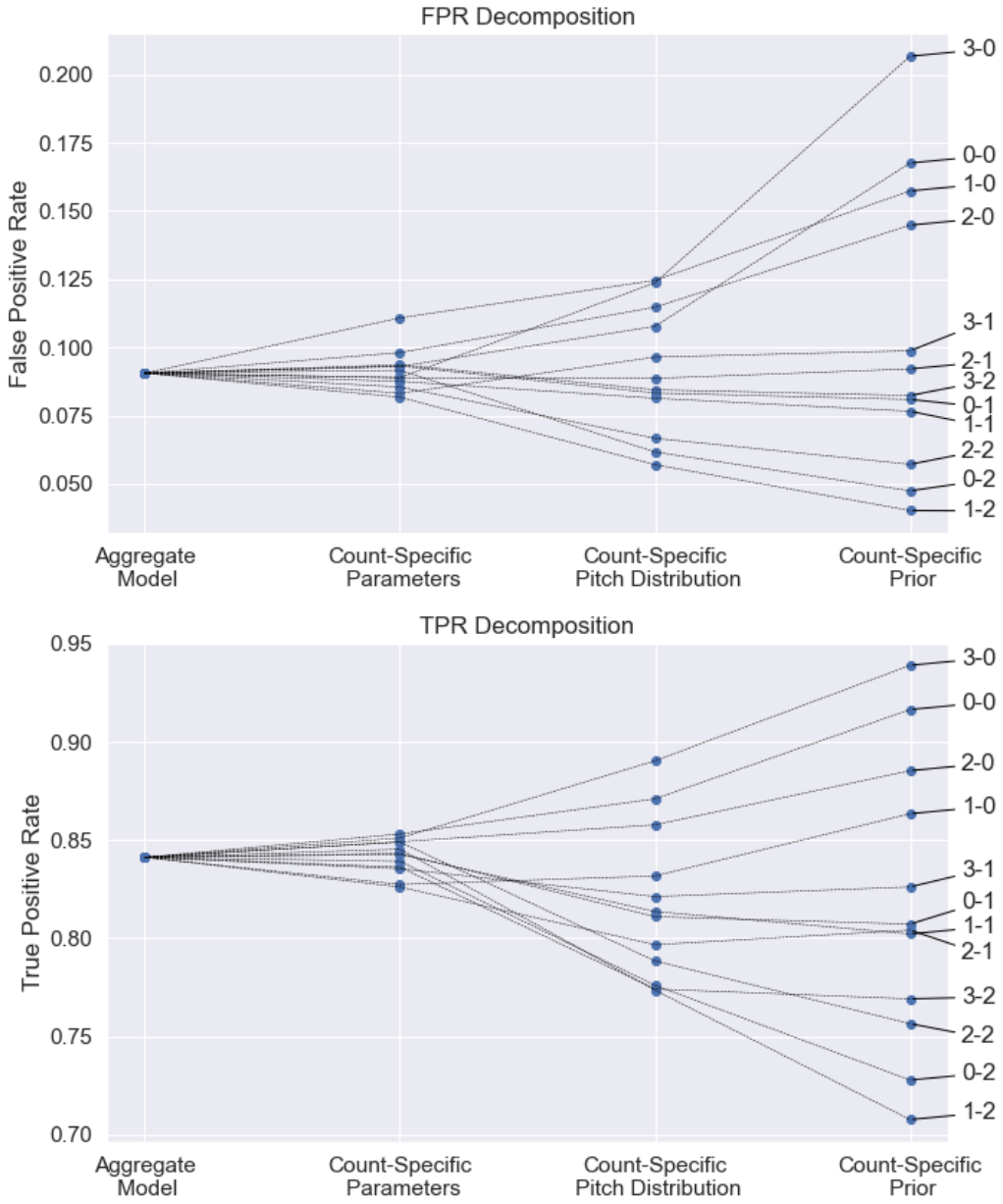
$$1 = \underbrace{\frac{TPR_4^c - TPR_3^c}{TPR_4^c - TPR_1^c}}_{\text{prior}} + \underbrace{\frac{TPR_3^c - TPR_2^c}{TPR_4^c - TPR_1^c}}_{\text{pitch location}} + \underbrace{\frac{TPR_2^c - TPR_1^c}{TPR_4^c - TPR_1^c}}_{\text{parameters}}$$

where TPR_m^c refers to the true positive rate of model m in count c. I use the analogous expression for false negatives.[24]

Averaging these decompositions across counts, I find that (parameters, pitch distribution, location) are responsible for (-14%, 57%, 57%) of the change in false positives and (4%, 59%, 36%) of the change in true positives.

[24]There are of course other ways to perform this decomposition. Since the model is non-linear, values will vary depending on the order in which sources of variation are added in. You could also use absolute values or squared differences. Regardless of specification, the general conclusion that all three sources are important contributors to observed behavior remains.

Figure 3.6: Decomposition of True and False Positive Rates by Source



Description: Predicted false positive and true positive rates from four versions of the model. Models vary on three dimensions:

- Aggregate Parameters (constant umpire parameters across counts) vs. Count-Specific Parameters.
- Aggregate Pitch Distribution (constant pitch distributions across counts) vs. Count-Specific Pitch Distributions.
- Aggregate Prior (the umpire uses the aggregate pitch distribution as his prior) vs. Count-Specific Priors.

Aggregate and count-specific values are constructed using 2019 data. Aggregate parameters are set to the mean of the estimated 2019 count-specific parameters.

In the four models above, I start with a model with aggregate preferences, pitch distribution, and prior (left). I then sequentially incorporate count-specific parameters, pitch distributions, and priors. The changes are additive: the final column (right) represents a model with count-specific parameters, pitch distributions, and priors.

Ignoring Observable Risk In Figure 3.7 I report the posterior means when $(\sigma_x, \sigma_z, \beta)$ are estimated using aggregate data. That is, I estimate a model in which both preferences and the umpire's prior are fixed across counts.

There are a few noticeable differences. First, the accuracy parameter σ_x is approximately 10-20% lower than I report in Figure 3.5. This is because discriminating by count allows umpires to increase their aggregate accuracy for fixed values of (σ_x, σ_z). When estimated on aggregate data, the model interprets that additional accuracy as increased precision on the part of umpires.

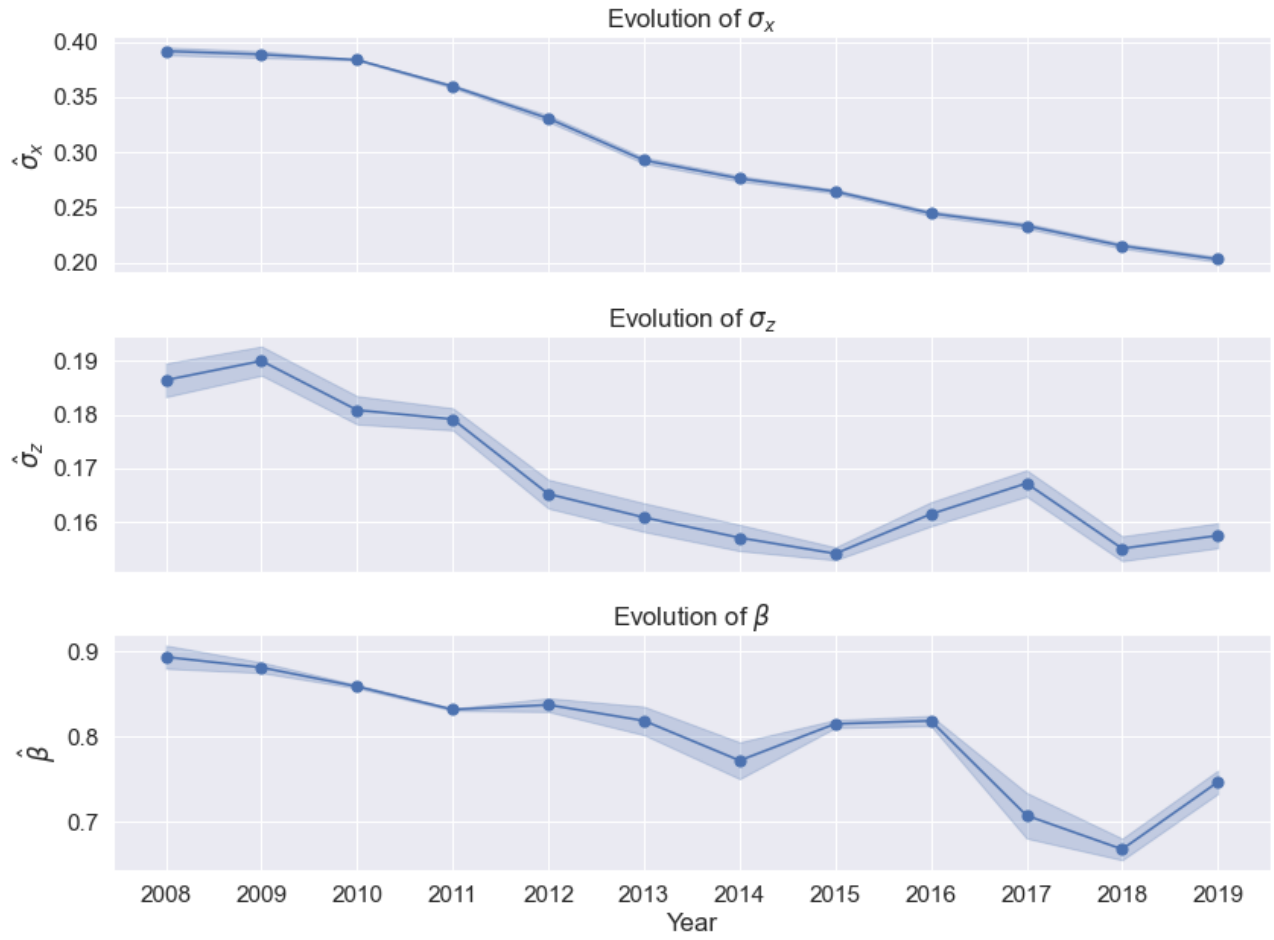
More interestingly, the aggregate model comes to a rather different conclusion regarding umpire preference parameters than the count-specific model. As opposed to finding false negatives roughly 20-40% more costly than a false positive, I now find false negatives to be roughly 10-30% *less* costly than a false positive.

The source of this discrepancy depends on the nature of count-specific information. In Figure 3.8, I plot the Hellinger distance between aggregate and pitch-specific count distributions in 2019.[25] The counts that are most distinct from the aggregate pitch distribution are 3-0, 0-2, 1-2, and, less dramatically, 2-2. The two-strike counts are ten times more common, accounting for 18.8% of all pitches versus 1.8% of pitches in 3-0 counts. In addition, umpires tend to shrink their strike zone in two-strike counts, lowering their false positive rate and increasing their false negative rate. The net effect of incorporating count-specific information, therefore, is towards higher false negative rates. A model that excludes count specific information interprets this tendency towards false negatives as evidence that false negatives are relatively cheap, i.e. that β is low.

[25]The Hellinger distance is a metric on probability distributions used to measure their similarity. For two discrete distributions p and q, their Hellinger distance is:

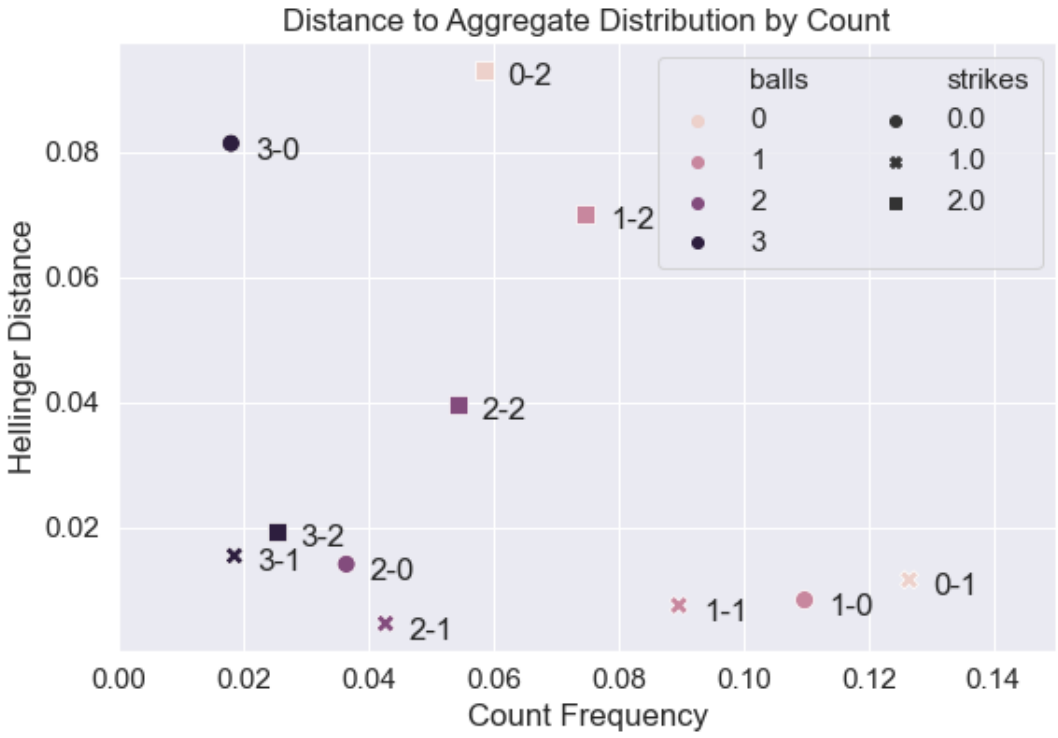
$$d(p, q) = \sum_i \left(\sqrt{p_i} - \sqrt{q_i}\right)^2 = \|\sqrt{p} - \sqrt{q}\|_2$$

Figure 3.7: Estimated Parameters Using Aggregated Data



Description: Posterior means of $(\sigma_x, \sigma_z, \beta)$ estimated using aggregate (not count-specific) data. The estimation procedure is identical to that for the count-specific model. Shaded areas represent plus/minus two estimated standard deviations.

Figure 3.8: Calculating Count-Specific Information



Description: This figure plots the Hellinger distance between the count-specific distribution of pitches and the aggregate distribution of pitches in 2019. To construct this measure, I estimate Gaussian kernel densities on a 30×30 strike zone grid with $x \in [-2.5, 2.5]$ and $z \in [0, 5]$. I use a bandwidth of 0.05, or 0.6 inches. For each point i on the grid I estimate the aggregate density a_i and the count-specific density c_i. The Hellinger distance is

$$d(a, c) = \sum_i \left(\sqrt{a_i} - \sqrt{c_i}\right)^2$$

The plot excludes the count 0-0 at (0.35, 0.016) for legibility purposes.

3.6 Conclusion

This chapter studies the ball/strike decisions made by Major League Baseball umpires. I find that previous claims of umpire bias can be given a simple rational explanation: umpires are Bayesian. They recognize that their signal of pitch location is noisy, and they err towards calling strikes in counts when pitchers almost always throw strikes. While I cannot rule out the possibility that umpire behavior is driven entirely by differences in preferences across counts, this offers a more parsimonious description of umpire behavior than previous literature (Walsh 2010, Green and Daniels 2014, D. Chen et al. 2016).

While this description implies substantial sophistication on the part of umpires, they have several factors working in their favor. First, classifying a pitch as a ball or strike is a significantly lower-dimensional problem than deciding whether to approve a prospective lender or job applicant. Second, to borrow an obvious metaphor, umpires get a lot of at-bats. A typical umpire calls more like 4,000 pitches a year and receives detailed feedback on every performance. Third, pitchers and batters exhibit easily-observed tendencies throughout the course of a game. These factors imply that umpires face a simple classification problem with well-known priors, conditions that may be conducive to smart decision-making. Nevertheless, given the extremely short timescales they operate on, it is impressive how closely umpire behavior follows a model of optimal decision-making.

References

Abito, Jose Miguel (n.d.). *Measuring the Welfare Gains from Optimal Incentive Regulation*. Working Paper. University of Pennsylvania (cit. on p. 3).

Arieff, Irwin (Oct. 2009). "A Witness Program for Elevators". In: *New York Times* (cit. on p. 10).

Armstrong, Mark and Sappington, David (2007). "Recent Developments in the Theory of Regulation". In: *Handbook of Industrial Organization*. Ed. by Mark Armstrong and Robert Porter. 1st ed. Vol. 3. Elsevier, pp. 1557–1700 (cit. on p. 7).

Arrow, Kenneth (1973). "The Theory of Discrimination". In: *Discrimination in Labor Markets*. Ed. by Orley Ashenfelter and Albert Rees. Princeton University Press (cit. on p. 96).

Association, National Restaurant (2019). *New York Restaurant Industry at a Glance* (cit. on p. 50).

Baker, George; Gibbons, Robert, and Murphy, Kevin (2002). "Relational Contracts and the Theory of the Firm". In: *The Quarterly Journal of Economics* 117 (1), pp. 39–84 (cit. on p. 8).

Bar-Isaac, Heski; Caruana, Guillermo, and Cuñat, Vincente (2012). "Information Gathering Externalities for a Multi-Attribute Good". In: *Journal of Industrial Economics* 60 (1), pp. 162–185 (cit. on pp. 1, 42).

Barber, Brad and Odean, Terrance (2001). "Boys will be Boys: Gender, Overconfidence, and Common Stock Investment". In: *The Quarterly Journal of Economics* 116 (1), pp. 261–292 (cit. on p. 94).

Becker, Gary (1957). *The Economics of Discrimination*. The University of Chicago Press (cit. on p. 96).

Bertrand, Marianne and Duflo, Esther (2016). *Field Experiments on Discrimination*. NBER Working Paper Series (cit. on p. 96).

Blackwell, David (1953). "Equivalent Comparisons of Experiments". In: *Annals of Mathematical Statistics* 24 (2), pp. 265–272 (cit. on p. 98).

Bloom, Nicholas; Sadun, Raffaella, and Reenen, John Van (2017). *Management as a Technology?* NBER Working Paper (cit. on p. 20).

Blundell, Wesley; Gowrisankaran, Gautam, and Langer, Ashley (2020). "Escalation of Scrutiny: The Gains from Dynamic Enforcement of Environmental Regulations". In: *American Economic Review* 110 (8), pp. 2558–2585 (cit. on pp. 5, 46).

Bohren, J. Aislinn et al. (2019). *Inaccurate Statistical Discrimination: An Identification Problem*. NBER Working Paper Series (cit. on pp. 97, 156).

Bolton, Patrick; Freixas, Xavier, and Shapiro, Joel (2012). "The Credit Ratings Game". In: *The Journal of Finance* 67 (1), pp. 85–111 (cit. on p. 42).

Cabral, Luís and Hortaçsu, Ali (2010). "The Dynamics of Seller Reputation: Evidence from eBay". In: *The Journal of Industrial Economics* 58 (1), pp. 54–78 (cit. on p. 1).

Chan, David; Gentzkow, Matthew, and Yu, Chuan (2019). *Selection with Variation in Diagnostic Skill: Evidence from Radiologists*. NBER Working Paper Series (cit. on pp. 69, 96–98, 101, 107).

Chen, Daniel; Moskowitz, Tobias, and Shue, Kelly (2016). "Decision Making Under the Gambler's Fallacy: Evidence from Asylum Judges, Loan Officers, and Baseball Umpires". In: *The Quarterly Journal of Economics* 131 (3), pp. 1181–1242 (cit. on pp. 95, 119).

Chen, Songnian and Khan, Shakeeb (2003). "Rates of convergence for estimation regression coefficients in heteroskedastic discrete response models." In: *Journal of Econometrics* 117.2, pp. 245–278 (cit. on pp. 143, 145).

Chiappori, P.A.; Levitt, S., and Groseclose, T. (2002). "Testing Mixed-Strategy Equilibria When Players Are Heterogeneous: The Case of Penalty Kicks in Soccer". In: *American Economic Review* 92 (4), pp. 1138–1151 (cit. on p. 94).

Dai, Weijia and Luca, Michael (2020). "Digitizing Disclosure: The Case of Restaurant Hygiene Scores." In: *American Economic Journal: Microeconomics* 12 (2), pp. 41–59 (cit. on pp. 42, 50, 56).

Darby, Michael and Karni, Edi (1973). "Free Competition and the Optimal Amount of Fraud". In: *The Journal of Law & Economics* 16 (1), pp. 67–88 (cit. on p. 6).

Davis, Noah and Lopez, Michael (2015). *Umpires Are Less Blind Than They Used To Be*. FiveThirtyEight. Last accessed December 13, 2020. URL available upon request. (Cit. on p. 110).

Dranove, David and Jin, Ginger (2010). "Quality Disclosure and Certification: Theory and Practice". In: *Journal of Economic Literature* 48 (4), pp. 935–963 (cit. on pp. 1, 2).

Dranove, David and Sfekas, Andrew (2008). "Start Spreading the News: A Structural Estimate of the Effects of New York Hospital Report Cards." In: *Journal of Health Economics* 27 (5), pp. 1201–1207 (cit. on pp. 1, 42).

Duflo, Esther et al. (2018). "The Value of Regulatory Discretion: Estimates from Environmental Inspections in India." In: *Econometrica* 86 (6), pp. 2123–2160 (cit. on p. 5).

Farhi, Emmanuel; Lerner, Josh, and Tirole, Jean (2013). "Fear of Rejection? Tiered Certification and Transparency." In: *RAND Journal of Economics* 44 (4), pp. 610–631 (cit. on p. 1).

Fast, Mike (2010). "What the Heck is PITCHf/x?" In: *The Hardball Times Baseball Annual* (cit. on p. 101).

— (2011). "Spinning Yarn: How Accurate is PitchTrax". In: *Baseball Prospectus* (cit. on p. 108).

Gagnepain, Philippe and Ivaldi, Marc (2002). "Incentive Regulatory Policies: The Case of Public Transit in France". In: *RAND Journal of Economics* 33 (4), pp. 605–629 (cit. on p. 3).

Galenianos, Manolis and Gavazza, Alessandro (2017). "A Structural Model of the Retail Market for Illicit Drugs." In: *American Economic Review* 107 (3), pp. 858–896 (cit. on pp. 4, 27).

Geman, Stuart and Hwang, Chii-Ruey (1982). "Nonparametric Maximum Likelihood Estimation by the Method of Sieves". In: *The Annals of Statistics* 10 (2), pp. 401–414 (cit. on p. 29).

Green, Etan and Daniels, David (2014). *What Does it Take to Call a Strike? Three Biases in Umpire Decision Making*. MIT Sloan Sports Analytics Conference (cit. on pp. 95, 103, 105, 119).

— (2018). *Bayesian Instinct* (cit. on p. 95).

Grossman, Sanford (1981). "The Informational Role of Warranties and Private Disclosure about Product Quality". In: *The Journal of Law & Economics* 24 (3), pp. 461–483 (cit. on p. 1).

Grossman, Sanford and Hart, Oliver (1980). "Disclosure Laws and Takeover Bids." In: *Journal of Finance* 35 (2), pp. 323–334 (cit. on p. 1).

Guderian, Doug (2014). *Elevator Maintenance Contracts 101*. Last accessed December 13, 2020. URL available upon request. (Cit. on p. 7).

Hastings, Justine and Weinstein, Jeffrey (2008). "Information, School Choice, and Academic Achievement: Evidence from Two Experiments". In: *The Quarterly Journal of Economics* 123 (4), pp. 1373–1414 (cit. on p. 1).

Ho, Daniel (2012). "Fudging the Nudge: Information Disclosure and Restaurant Grading". In: *The Yale Law Journal* 122 (3), pp. 574–688 (cit. on pp. 44, 53, 55).

Holmström, Bength (1999). "Managerial Incentive Problems — A Dynamic Perspective". In: *Review of Economic Studies* 66 (1), pp. 169–182 (cit. on p. 20).

Jacob, Brian and Levitt, Steven (2003). "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating." In: *Quarterly Journal of Economics* 118 (3), pp. 843–877 (cit. on pp. 1, 42).

Jin, Ginger and Leslie, Phillip (2003). "The Effect of Information on Product Quality: Evidence from Restaurant Hygiene Grade Cards." In: *The Quarterly Journal of Economics* 118 (2), pp. 409–451 (cit. on pp. 1, 44, 50, 53, 55, 79).

Jin, Ginger and Sorensen, Alan (2006). "Information and Consumer Choice: The Value of Publicized Health Plan Ratings." In: *Journal of Health Economics* 25 (2), pp. 248–275 (cit. on p. 1).

Jovanovic, Boyan (1982). "Truthful Disclosure of Information." In: *Bell Journal of Economics* 13 (1), pp. 36–44 (cit. on p. 1).

Kaggle (2015). *Elevators in New York City*. Last accessed December 13, 2020. URL available upon request. (Cit. on p. 11).

Kang, Karam and Silveira, Bernardo (2020). *Understanding Disparities in Punishment: Regulator Preferences and Expertise*. Working Paper (cit. on p. 5).

Knowles, John; Persico, Nicola, and Todd, Petra (2001). "Racial Bias in Motor Vehicle Searches: Theory and Evidence". In: *Journal of Political Economy* 109 (1), pp. 203–229 (cit. on p. 97).

Krishna, Priya (June 2018). "The Life of a Restaurant Inspector: Rising Grades, Fainting Owners". In: *New York Times* (cit. on p. 47).

Kroll, Karen (2017). *The Right Elevator Service Agreement Minimizes Downtime*. Last accessed December 13, 2020. URL available upon request. (Cit. on p. 7).

Leaver, Clare (2009). "Bureaucratic Minimal Squawk Behavior: Theory and Evidence from Regulatory Agencies." In: *American Economic Review* 99 (3), pp. 572–607 (cit. on pp. 46, 65).

Lewbel, Arthur (2000). "Semiparametric qualitative response model estimation with unknown heteroscedasticity or instrumental variables." In: *Journal of Econometrics* 97 (1), pp. 145–177 (cit. on p. 143).

Lim, Claire and Yurukoglu, Ali (2018). "Dynamic Natural Monopoly Regulation: Time Inconsistency, Moral Hazard, and Political Environments". In: *Journal of Political Economy* 126 (1), pp. 263–312 (cit. on p. 3).

Lowe, Luther (2018). *Bringing Restaurant Hygiene Scores to Yelp Pages Everywhere*. Yelp. Last accessed December 13, 2020. URL available upon request. (Cit. on p. 50).

Lu, Susan (2012). "Multitasking, Information Disclosure, and Product Quality: Evidence from Nursing Homes". In: *Journal of Economics & Management Strategy* 21 (3), pp. 673–705 (cit. on p. 42).

Mailath, George and Samuelson, Larry (2001). "Who Wants a Good Reputation?" In: *Review of Economic Studies* 68 (2), pp. 415–441 (cit. on p. 20).

Makofske, Matthew (2020a). "Mandatory disclosure, letter-grade systems, and corruption: The case of Los Angeles County restaurant inspections." In: *Journal of Economic Behavior & Organization* 172, pp. 292–313 (cit. on pp. 42–44, 55, 57).

— (2020b). "The Effect of Information Salience on Product Quality: Louisville Restaurant Hygiene and Yelp.com". In: *Journal of Industrial Economics* 68 (1), pp. 52–92 (cit. on pp. 42, 50).

McCann, Michael (2013). *Deaths and Injuries Involving Elevators and Escalators*. Research Report. The Center for Construction Research and Training (cit. on p. 6).

McCormick PCS (n.d.). *Elevator Maintenance Agreement Specs*. Last accessed December 13, 2020. URL or document available upon request. (Cit. on p. 6).

Milgrom, Paul (1981). "Good News and Bad News: Representation Theorems and Applications." In: *Bell Journal of Economics* 12 (2), pp. 380–391 (cit. on p. 1).

Mills, Brian (2014). "Social Pressure at the Plate: Inequality Aversion, Status, and Mere Exposure". In: *Managerial and Decision Economics* 35 (6), pp. 387–403 (cit. on p. 95).

MLB Playing Rules Committee (2019). *Official Baseball Rules*. 2019 Edition (cit. on pp. 100, 102).

Nickerson, Raymond (1998). "Confirmation Bias: A Ubiquitous Phenomenon in Many Guises". In: *Review of General Psychology* 2 (2), pp. 175–220 (cit. on p. 94).

NYC Department of Buildings (n.d.). *Department of Buildings Guide To: Elevators*. Last accessed December 13, 2020. URL available upon request. (Cit. on p. 6).

NYC Department of Investigation (June 2008). *DOI Arrests City Restaurant Inspector in Bribery "Sting"* (cit. on p. 63).

NYC Department of Mental Health and Hygiene (Dec. 2010). *Chapter 23 of Title 24 of the Rules of the City of New York. Inspection Scoring and Letter Grading System for Food Service Establishments* (cit. on p. 49).

— (Oct. 2011). *Requirements for Posting Letter Grade Cards* (cit. on p. 47).

— (July 2014). *Changes to Inspection Fines for Restaurants: What You Need to Know* (cit. on p. 49).

— (June 2016). *Self-Inspection Worksheet for Food Service Establishments* (cit. on p. 55).

NYC Open Data (2020a). *311 Service Requests from 2010 to Present*. Last accessed December 13, 2020. URL available upon request. (Cit. on p. 53).

— (2020b). *DOB Job Application Filings*. Last accessed December 13, 2020. URL available upon request. (Cit. on p. 11).

— (2020c). *DOHMH New York City Restaurant Inspection Results*. Last accessed December 13, 2020. URL available upon request. (Cit. on p. 51).

Phelps, Edmund (1972). "The Statistical Theory of Racism and Sexism". In: *The American Economic Review* 62 (4), pp. 659–661 (cit. on p. 96).

Pope, Devin and Sydnor, Juston (2011). "What's in a Picture? Evidence of Discrimination from Prosper.com". In: *The Journal of Human Resources* 46 (1), pp. 53–92 (cit. on p. 97).

Rothbart, Michah et al. (2014). *The Impact of Restaurant Letter Grades on Taxes and Sales: Micro Evidence from New York City*. Annual Conference on Taxation and Minutes of the Annual Meeting of the National Tax Association, pp. 1–44 (cit. on p. 50).

Rust, John (1987). "Optimal Replacement of GMC Bus Engines: An Empirical Model of Harold Zurcher". In: *Econometrica* 55 (5), pp. 999–1033 (cit. on p. 6).

Ryan, Stephen (2012). "The Costs of Environmental Regulation in a Concentrated Industry". In: *Econometrica* 80 (3), pp. 1019–1061 (cit. on p. 3).

Scallan, Elaine; Griffin, Patricia, et al. (2011). "Foodborne Illness Acquired in the United States—Unspecified Agents". In: *Emerging Infectious Diseases* 17 (1), pp. 16–22 (cit. on p. 52).

Scallan, Elaine; Hoekstra, Robert, et al. (2011). "Foodborne Illness Acquired in the United States—Major Pathogens". In: *Emerging Infectious Diseases* 17 (1), pp. 7–15 (cit. on p. 52).

Schifman, Gerald (Mar. 2018). "The Lurking Error in Statcast Pitch Data". In: *The Hardball Times* (cit. on p. 101).

Schneider, Henry (2012). "Agency Problems and Reputation in Expert Services: Evidence from Auto Repair". In: *Journal of Industrial Economics* 60.3, pp. 406–433 (cit. on p. 6).

Source, Elevator (n.d.). *Elevator Life Expectancy*. Last accessed December 13, 2020. URL available upon request. (Cit. on p. 6).

Stigler, George (1971). "The Theory of Economic Regulation". In: *The Bell Journal of Economics and Management Science* 2 (1), pp. 3–21 (cit. on p. 5).

Timmins, Christopher (2002). "Measuring the Dynamic Efficiency Costs of Regulators' Preferences: Municipal Water Utilities in the Arid West". In: *Econometrica* 70 (2), pp. 603–629 (cit. on p. 3).

Tversky, Amos and Kahneman, Daniel (1974). "Judgement under Uncertainty: Heuristics and Biases". In: *Science* 185 (4157), pp. 1124–1131 (cit. on p. 94).

Walsh, John (2010). *The Compassionate Umpire*. (Visited on 11/04/2020) (cit. on pp. 95, 119).

Werner, Rachel and Asch, David (2005). "The Unintended Consequences of Publicly Reporting Quality Information." In: *Journal of the American Medical Association* 293 (10), pp. 1239–1244 (cit. on p. 1).

Willett-Wei, Megan (Mar. 2014). "New York City's Restaurant Grading System Is In Need Of A Major Overhaul". In: *Business Insider* (cit. on p. 86).

Wolak, Frank (1994). "An Econometric Analysis of the Asymmetric Information, Regulator-Utility Interaction". In: *Annals of Economics and Statistics* (34), pp. 1–12 (cit. on p. 3).

Yinger, John (1996). "Why Default Rates Cannot Shed Light on Mortgage Discrimination". In: *Cityscape: A Journal of Policy Development and Research* 2 (1), pp. 25–32 (cit. on p. 97).

Appendix A: Appendix For "Regulation by Information Provision'

A.1 Quality Upgrading

This section argues that buildings tend to switch to higher-quality service providers in response to DOB violations. To do this, I calculate each service provider's yearly violations per machines. For every year in which a building cancels a contract, I then compare the violation rate of their new provider to that of their previous provider. As shown in Table A.1, when an ECB violation was present in year $t-1$, the average change in violation rates is 1.26 percentage points small vs. when no ECB violation was present. That is, if the average change in violation rates were 0 (as we would expect in a steady state), then cancellations that took place following an ECB violation would tend to be towards higher quality providers.

A.2 Quality Persistence

This section tries to determine whether service providers vary their effort over the length of contracts. Obviously regressing a building's violation incidence in year t against their contract length in year t suffers from omitted variable bias: long-lived contracts will tend to be with high-performing providers, inducing a natural negative correlation between these two measures. In Table A.2, I also include fixed effects for the total length of a building's contract with their current provider. As expected, the coefficient on total contract length is negative: the longer-lasting contracts have fewer violations.

Of more interest is the coefficient on current contract length, which — while statistically significant — is small in magnitude: it suggests that for a 10-year contract, the provider will generate about 0.013 more violations per year in the final year of the contract relative to the initial year (the average building accrues 0.197 violations per year in this dataset). This suggests that providers

Table A.1: Quality Upgrading Regression

	coef	std err	t	P>|t|
$\mathbb{I}\{\text{ECB Violations in } t-1\}$	-1.26	0.16	-7.85	0.000
R^2			0.148	
N			36,446	
Fixed Effects			Year, Ownership, Occupancy	
Other controls			Building Height, No. Machines, Contract Length	

Description: Unit of observation: building-year. Sample: all building-years in which a building switched providers. Dependent variable: difference in service provider violations/machine (across all buildings) from $t-1$ to t. Units are from 0-100, so regression coefficients indicate percentage point differences in violation rates.

are not changing their effort dramatically throughout the course of their contracts, in line with the fixed effort assumption embedded in the model of Section 1.4.

A.3 Descriptive Data

A.3.1 Service Provider Concentration and Entry

The theoretical model presented in Section 1.4 considers a mass of service providers and does not allow for provider entry. Table A.3 provides descriptive evidence to support these choices. The mapping of service providers to buildings lets me track entry, exit, and provider market share across time. The service industry is quite fragmented, with 80-90 providers active in most years and a Herfindahl index less than 250. Turnover is also limited, with approximately 3 new entrants per year, gaining on average 1% market share in the year of entry.

A.3.2 Missed Inspection Patterns

The key exclusion restriction for the IV analysis in Section 1.3 is that missed routine inspections are uncorrelated with factors involved in a building's decision to retain their service provider. While institutional details support this restriction — the DOB does not have a stated policy of directing their resources towards particular buildings or service providers — it also appears to be

Table A.2: Quality Persistence Regression

	coef	std err	t	P>|t|
Intercept	0.245	0.062	3.959	0.000
Current Contract Length	0.0013	0.001	2.109	0.035
Total Contract Length (excl = 1)				
2	0.0460	0.016	2.923	0.003
3	-0.0482	0.015	-3.222	0.001
4	-0.0323	0.015	-2.165	0.030
5	-0.0415	0.015	-2.814	0.005
[Contract lengths above 5 excluded from table]				
R^2	0.03			
N	124,302			
Fixed Effects	Year, Total Contract Length, Ownership, Occupancy			
Other controls	No. Machines			

Description: Unit of observation: building-year. Sample: all building-years in Manhattan since 2000. Dependent variable: number of ECB violations.

the case in the data.

Table A.4 estimates the probability that a given machine will be visited in a given year as a function of its building's characteristics, time, and the building's service provider. While most of the coefficients are statistically significant, few are economically meaningful. The difference in inspection rates between residential and commercial buildings - as well as between corporate and non-corporately owned ones - is approximately one percentage point. There may be a geographic bias against the Bronx, where inspections are performed around two points less frequently. Time variation in enforcement is the most useful from an explanatory perspective, which makes sense given the variation in inspection rates already mentioned. Lastly, service provider fixed effects add very little to the explanatory power of the model. This all suggests that random inspections is a good approximation of the regulator's current policy.

Table A.3: Service Provider Concentration

	1996 - 2007	2008 - 2016
Market Concentration		
Avg. Active Providers	88.0	81.2
Herfindahl Index (0-10,000)	216.1	249.7
Mkt. Share of Top Four Providers	13.1	16.2
% Active in >= 4 Boroughs	79.7	84.1
Entry and Exit		
Avg. New Providers	3.0	2.9
Mkt. Share of New (%)	1.6	0.6
Avg. Exiting Providers	4.2	2.8
Mkt. Share of Exit (%)	2.8	1.1

Description: Descriptive statistics of service provider concentration and entry. Providers must work in at least 10 buildings in a year to be considered in the market, and 250 throughout the panel to be part of the sample. These cutoffs capture over 99% of inspections.

A.4 Model Proofs

A.4.1 Proof of Proposition 1.4.1

Proposition. *(1.4.1) For distributions $F(x), G(x)$, let $F \succ G$ denote that F first-order stochastically dominates G. The building's posterior distribution satisfies*

$$P(\theta|\hat{e}, \kappa, n+1, v) \succ P(\theta|\hat{e}, \kappa, n, v) \succ P(\theta|\hat{e}, \kappa, n, v+1) \tag{A.1}$$

Proof. For convenience, write $e = \hat{e}\kappa$ and drop the $\hat{e}$ and κ dependence. I will show $P(\theta|n+1, v) \succ P(\theta|n, v)$; the proof that $P(\theta|n, v) \succ P(\theta|n, v+1)$ is nearly identical.

Subtracting and simplifying gives

$$P(\theta|n+1, v) - P(\theta|n, v) = \theta^n(1-e\theta)^v P(\theta) \left[\frac{\theta}{\int_0^1 \theta^{n+1}(1-e\theta)^v P(\theta)d\theta} - \frac{1}{\int_0^1 \theta^n(1-e\theta)^v P(\theta)d\theta} \right]$$

The quantity in brackets is strictly negative at $\theta = 0$, strictly positive at $\theta = 1$, and strictly increasing in between. Therefore, by the full support assumption, the difference in cumulative distributions

Table A.4: Regulatory Inspection Patterns

	100*P(Completed Routine Inspection)			
No. Machines	0.05***	0.05***	0.04***	0.03***
	(0.003)	(0.003)	(0.003)	(0.003)
Owner (Excl. = Corporate)				
Non-Corporate	-1.24***	-1.29***	-0.93***	-0.86***
	(0.08)	(0.07)	(0.08)	(0.07)
Gov't	-4.36***	-4.97***	-4.00***	-4.68***
	(0.24)	(0.23)	(0.25)	(0.24)
Occupant (Excl. = Commercial)				
Residential	-0.82***	-0.77***	-0.94***	-0.87***
	(0.09)	(0.09)	(0.09)	(0.09)
Other	-4.54***	-4.90***	-4.43***	-4.59***
	(0.13)	(0.12)	(0.13)	(0.12)
Borough (Excl. = Bronx)				
Brooklyn	-2.40***	-2.59***	-2.64***	-2.20***
	(0.175)	(0.166)	(0.184)	(0.176)
Manhattan	-1.75***	-1.61***	-2.41***	-2.38***
	(0.15)	(0.14)	(0.16)	(0.15)
Queens	0.60***	0.55***	0.07	0.21
	(0.18)	(0.17)	(0.19)	(0.18)
Staten Island	-1.72***	-2.17***	-2.20***	-1.81***
	(0.38)	(0.36)	(0.39)	(0.37)
Year FE	N	Y	N	Y
SP FE	N	N	Y	Y
N	958,698			
R^2	0.00	0.11	0.02	0.11

Description: LPM of a completed routine inspection (excluding visits where the inspector could not gain access to the machine). Unit of observation is a machine-year, and only includes years machines are active. Units are from 0-100, so regression coefficients indicate percentage point differences in inspection rates.

$F(\theta|n+1,v) - F(\theta|n,v)$ is strictly decreasing on some interval $[0,\theta^*)$ and strictly increasing on $(\theta^*,1]$. Since $F(\theta|n+1,v) - F(\theta|n,v)$ is 0 at $\theta = 0$ and 1,

$$F(\theta|n+1,v) \leq F(\theta|n,v)\ \forall\theta \in [0,1] \tag{A.2}$$

That is, $F(\theta|n+1,v)$ dominates $F(\theta|n,v)$.

□

A.4.2 Proof of Proposition 1.4.2

Proposition. *(1.4.2) For any outside option $\bar{V}$ and effort beliefs $\hat{e}_s$, a unique value function $V_s(\mathcal{I},\epsilon,\hat{e}_s,\bar{V})$ exists. The building's optimal policy is a cutoff rule such that offers are accepted if and only if $\epsilon \geq \epsilon_s^*(\mathcal{I},\hat{e}_s,\bar{V})$. The cutoffs are characterized by the indifference condition*

$$\bar{V}(1-\beta) = \alpha P_s(n|\mathcal{I},\hat{e}_s) - p_s + \sigma\epsilon_s^*(\mathcal{I}) + \sigma\beta\mathbb{E}_s\left[\max\{\epsilon - \epsilon_s^*(\mathcal{I}'),0\}\right] \tag{A.3}$$

where $\mathbb{E}_s$ is taken over the distribution of ϵ and this period's regulatory outcomes.

Moreover, the value function V_s and the cutoff ϵ_s^* are:

- *Weakly decreasing (strictly increasing) in price*
- *Weakly increasing (strictly decreasing) in good signals*
- *Weakly decreasing (strictly increasing) in bad signals*

Proof. For convenience I drop provider subscripts s throughout the proof.

Let $\mathcal{B}$ be the set of bounded functions from $\mathbb{N}^2$ to $\mathbb{R}$. Define a *threshold function* $f \in \mathcal{B}$ and its associated value function W_f as

$$W_f(\mathcal{I},\epsilon) = \begin{cases} \bar{V} & \epsilon \leq f(\mathcal{I}) \\ \bar{V} + \sigma(\epsilon - f(\mathcal{I})) & \epsilon \geq f(\mathcal{I}) \end{cases}$$

For any threshold functions f, g:

$$\sup_{\mathcal{I},\epsilon}|W_f(\mathcal{I},\epsilon) - W_g(\mathcal{I},\epsilon)| = \sigma \sup_{\mathcal{I}}|f(\mathcal{I}) - g(\mathcal{I})| = \sigma d_\infty(f,g)$$

By Assumption 1.4.2, the expected value of the value function is bounded:

$$\begin{aligned}\left|\int_\epsilon W(\mathcal{I},\epsilon)dF(\epsilon)\right| &= \left|\bar{V} + \sigma\int_{f(\mathcal{I})}^{\infty}(\epsilon - f(\mathcal{I}))dF(\epsilon)\right| \\ &\leq |\bar{V}| + \sigma\int_{-\infty}^{\infty}|\epsilon| + \sigma|f(\mathcal{I})|\end{aligned}$$

Define the Bellman operator

$$T(W_f)(\mathcal{I},\epsilon) = \max\left\{\alpha P(n|\mathcal{I}) - p + \sigma\epsilon + \beta\mathbb{E}(W_f(\mathcal{I}',\epsilon)|\mathcal{I}), \bar{V}\right\}$$

Note $T(W_f) = W_{T(f)}$ where $T(f)$ is defined by:

$$\alpha P(n|\mathcal{I}) - p + \sigma T(f)(\mathcal{I}) + \beta\mathbb{E}(W_f(\mathcal{I}',\epsilon)|\mathcal{I}) = \bar{V} \tag{A.4}$$

The fact that $T(f)$ is bounded follows from the boundedness of $\int_\epsilon W_f$ and $P(n|\mathcal{I})$. Combining Equation A.4 with the fact that $W_f(\mathcal{I},\epsilon') - W_f(\mathcal{I},\epsilon) = \sigma(\epsilon' - \epsilon)$ implies the indifference condition noted in the statement of the proposition.

Differencing Equation A.4 for threshold functions f, g gives:

$$\sigma\,|T(f)(\mathcal{I}) - T(g)(\mathcal{I})| = \beta\left|\mathbb{E}(W_f(\mathcal{I}',\epsilon) - W_g(\mathcal{I}',\epsilon)|\mathcal{I})\right| \leq \sigma\beta d_\infty(f,g)$$

Taking the sup shows that T induces a contraction on the space of threshold functions, which is a complete metric space. Thus there is a unique value function and policy obtained by iterating T.

Comparative statics follow from standard dynamic programming arguments. Suppose W is

weakly decreasing in p. Then $T(W)$ is weakly decreasing as well. Since weak monotonicity is preserved in the limit, the value function V is weakly decreasing as well.[1]

Since the one-period payoff function is strictly monotone in p, n, and v (see Proposition 1.4.1), strict monotonicity of the policy function follow from the indifference condition A.4 and the weak monotonicity of the value function. □

A.4.3 Calculating the length of provider contracts

Proposition A.4.1. *Given building beliefs $\hat{e}_s$ and an outside option $\bar{V}$, the expected length of provider s's contract at information set $\mathcal{I}$ given unknown type θ can be found by iterating the recursive formula:*

$$\ell_s(\mathcal{I}) = d_s(\mathcal{I})\left[1 + re\kappa_s\theta\ell_s(\mathcal{I}_n) + r(1 - e\kappa_s\theta)\ell_s(\mathcal{I}_v) + (1 - r)\ell_s(\mathcal{I})\right] \tag{A.5}$$

where $d_s(\mathcal{I}) = 1 - F_\epsilon(\epsilon_s^*(\mathcal{I}), \hat{e}_s, \bar{V})$ is the probability s's offer is accepted at $\mathcal{I}$.[2] *Moreover, the expected length of a contract is*

- *Strictly increasing in e and n*
- *Strictly decreasing in p and v*

Proof. Let

$$d_s(p|\mathcal{I}, \hat{e}_s, \bar{V}) = 1 - F_\epsilon(\epsilon_s^*(\mathcal{I}, \hat{e}_s, \bar{V})) \tag{A.6}$$

denote the probability a contract is renewed conditional on reaching information set $\mathcal{I}$. Note d_s does not depend on provider effort, since they take the building's policy as fixed.

Let $\ell_s(p, e, \mathcal{I}, \theta|\hat{e}_s, \bar{V})$ denote the expected length of a contract conditional on being type θ and making an offer at information set $\mathcal{I}$. Expected length admits a recursive representation (suppress-

[1] The argument for n and v is a little more subtle since the relative probability of signals are changing as well. However, since an increase in n raises the probability of landing in the good state tomorrow, the result follows quickly.

[2] I have suppressed the dependence of ℓ_s on $p, e, \hat{e}_s$ and $\bar{V}$ for readability. $\mathcal{I}_n$ ($\mathcal{I}_v$) denotes the information set following a no-violation (in-violation) outcome.

ing the $p, e, \theta, \hat{e}_s$ and $\bar{V}$ dependence):

$$\ell_s(\mathcal{I}) = d_s(\mathcal{I}) \left[1 + re\kappa_s\theta\ell_s(\mathcal{I}_n) + r(1 - e\kappa_s\theta)\ell_s(\mathcal{I}_v) + (1 - r)\ell_s(\mathcal{I})\right] \tag{A.7}$$

Rearranging gives

$$\ell_s(\mathcal{I}) = \frac{d_s(\mathcal{I})}{1 - (1 - r)d_s(\mathcal{I})} \left[1 + re\kappa_s\theta\ell_s(\mathcal{I}_n) + r(1 - e\kappa_s\theta)\ell_s(\mathcal{I}_v)\right] \tag{A.8}$$

Given a guess ℓ_0, Equation A.8 defines an updating rule which it is immediate to show satisfies Blackwell's sufficient conditions, meaning ℓ_s is uniquely determined by iterating the updating rule.

Suppose ℓ_0 is increasing in n and e and decreasing in v and p. Since $d_s(\mathcal{I})$ is increasing in n and decreasing in v and p (see Proposition 1.4.2), Equation A.8 immediately shows that the updated guess, ℓ_1, obeys the same comparative statics with respect to e, n, and v. Therefore in the limit, ℓ_s must obey weak monotonicity in these arguments.

To show that ℓ_s is strictly increasing in n, first recall that $d_s(\mathcal{I})$ is strictly increasing in n by Proposition 1.4.1. Therefore:

$$\begin{aligned}
\ell_s(\mathcal{I}_n) &> \frac{d_s(\mathcal{I}_v)}{1 - (1 - r)d_s(\mathcal{I}_v)} \left[1 + re\kappa_s\theta\ell_s(p, e, \mathcal{I}_{nn}) + r(1 - e\kappa_s\theta)\ell_s(p, e, \mathcal{I}_{nv})\right] && \text{(A.9)} \\
&\geq \frac{d_s(\mathcal{I}_v)}{1 - (1 - r)d_s(\mathcal{I}_v)} \left[1 + re\kappa_s\theta\ell_s(p, e, \mathcal{I}_{vn}) + r(1 - e\kappa_s\theta)\ell_s(p, e, \mathcal{I}_{vv})\right] && \text{(A.10)} \\
&= \ell_s(\mathcal{I}_v) && \text{(A.11)}
\end{aligned}$$

The same technique shows that ℓ_s is strictly decreasing in v.

Lastly, to show ℓ_s is strictly increasing in e, simply differentiate Equation A.8 with respect to e. The derivative is proportional to $\ell_s(\mathcal{I}_n) - \ell_s(\mathcal{I}_v)$, which is strictly positive.

□

A.4.4 Proof of Proposition 1.4.3

I use following lemma to prove Proposition 1.4.3:

Lemma A.4.1. *Let x be a random variable with CDF F satisfying $\mathbb{E}(x) < \infty$. Then*

$$\lim_{x\to\infty} x(1 - F(x)) = 0 \tag{A.12}$$

Proof. If the limit is not 0, then there exists an $\epsilon > 0$ such that for any X, there exists $x > X$ satisfying $x(1 - F(x)) > \epsilon$.

Define a sequence $x_1, x_2, \ldots$ such that $x_n(1 - F(x_n)) > \epsilon$ and $x_{n+1} > 2x_n$. Note F first-order dominates the distribution defined by

$$\tilde{F}(x) = \begin{cases} F(x) & x \leq x_1 \\ 1 - \frac{\epsilon}{x_{n+1}} & x \in [x_n, x_{n+1}) \end{cases} \tag{A.13}$$

However, $\tilde{F}$ does not have a finite mean, since for $n \geq 2$, x_n contributes $\epsilon - \frac{\epsilon x_n}{x_{n+1}} > \epsilon/2$ to the mean, so the mean diverges. Since F dominates $\tilde{F}$, this implies F does not have a finite mean, a contradiction.

□

I now prove Proposition 1.4.3.

Proposition. *For any $\tilde{v}$ and any beliefs $\hat{e}_s$, $p_s^*(\hat{e}_s, \bar{V})$ and $e_s^*(\hat{e}, \bar{V})$ exist and are characterized by the first-order conditions of Equation 1.3.*

Proof. In the proof of Proposition A.4.1, I showed that for a fixed p, the threshold function $\epsilon^*(p, \mathcal{I}, \hat{e}_s, \bar{V})$ is bounded. Define $\bar{\epsilon}(p, \hat{e}_s, \bar{V}) = \lim_{g\to\infty} \epsilon^*(p, \{n, 0\}, \hat{e}_s, \bar{V})$. This limit exists because it is a decreasing sequence which is bounded below. From the monotonicity results in Proposition 1.4.2, $\bar{\epsilon}(p, \hat{e}_s, \bar{V}) < \epsilon^*(p, \mathcal{I}, \hat{e}_s, \bar{V})$ for all $\mathcal{I}$. Define

$$\bar{d}(p, \hat{e}_s, \bar{V}) = 1 - F_\epsilon(\bar{\epsilon}(p, \hat{e}_s, \bar{V})),$$

which is strictly less than 1 by the full-support assumption (Assumption 1.4.2). Since the proba-

bility of retaining a contract is less than $\bar{d}$ in every period, the provider's profits are bounded:

$$\pi_s(p, e|\theta)\ell_s(p, e|\theta, \hat{e}_s, \bar{V}) < p\frac{\bar{d}(p, \hat{e}_s, \bar{V})}{1 - \bar{d}(p, \hat{e}_s, \bar{V})}$$

Note Equation A.4 can be written as:

$$\bar{V}(1-\beta) = \alpha P_s(n|\mathcal{I}) - p + \sigma\epsilon_s^*(\mathcal{I}) + \sigma\beta \sum_{m\in\{n,v,e\}} P_s(m|\mathcal{I}) \left[\int_{\epsilon_s^*(\mathcal{I}_m)}^{\infty} (\epsilon - \epsilon_s^*(\mathcal{I}_m))dF(\epsilon)\right] \quad \text{(A.14)}$$

where n, v, e refer to a no-violation, in-violation, and empty regulatory message. Differentiating with respect to p shows:

$$\frac{\partial\epsilon_s^*(p, \mathcal{I}, \hat{e}_s, \bar{V})}{\partial p} > \sigma^{-1}, \quad \text{(A.15)}$$

so ϵ_s^* grows faster than linear in p. Combined with Lemma A.4.1, the limiting behavior of π_s is given by:

$$\lim_{p\to\infty} \pi_s(p, e|\theta)\ell_s(p, e|\theta, \hat{e}_s, \bar{V}) < p\frac{\bar{d}(p, \hat{e}_s, \bar{V})}{1 - \bar{d}(p, \hat{e}_s, \bar{V})} = 0 \quad \text{(A.16)}$$

Since it is possible for providers to generate a nonzero return in this model, there is a uniform upper bound the optimal p. Since effort is constrained to lie within the compact set $[0, \bar{e}]$, the provider's problem simplifies to maximizing a continuous objective over a compact set, meaning a maximum exists.

Finally, to show that the solution is characterized by a first-order condition, note that no point along the boundary is optimal. Clearly choosing $p = 0$ is suboptimal, since the provider earns a negative return in that case. At $e = 0$, profits are actually increasing in effort:

$$\left.\frac{\partial_e\pi_s(p, e|\theta)}{\partial e}\right|_{e=0} = -c'(0) + r\phi = r\phi > 0$$

Since share is also increasing in e, choosing $e = 0$ cannot be optimal.

Since cost diverges as e approaches $\bar{e}$, choosing $e = \bar{e}$ cannot be optimal either. Therefore the optimal price/effort combination is interior, and is characterized by a first-order condition.

□

A.5 Identification

A.5.1 Identification of the Cost Function

I first show that, for a provider with known prices, every term of Equation 1.6 is identified after multiplying by e (other than $c(e)$ and $ec'(e)$). The regulatory parameters r, ϕ are known, as are $e_s\kappa_s$ and the distribution of θ. Therefore the only terms remaining are ℓ_s and $e\partial_e\ell_s$. The contract length function ℓ_s depends only on the aggregate quantity $e_s\kappa_s$ (see Equation A.5). Therefore:

$$\ell_s(p, e|\theta, \hat{e}_s, \bar{V}) = f_s(p, e\kappa_s|\theta, \hat{e}_s, \bar{V}) \tag{A.17}$$

From the building's policy, f_s and its derivatives can all be calculated.[3] Differentiating and multiplying by e_s gives

$$e\partial_2\ell_s(p, e|\theta, \hat{e}_s, \bar{V}) = e\kappa_s\partial_2 f_s(p, e\kappa_s|\theta, \hat{e}_s, \bar{V}) \tag{A.18}$$

Since $e_s\kappa_s$ is identified and the partial of f can be calculated, the right-hand side is known.

A Simple Example

Consider the following problem which is closely related to the provider's optimization problem:

$$\max_{p,e}(p - ce + er)\ell(e, p)$$

where ℓ is a contract length function satisfying $\ell_s > 0$, $\ell_p < 0$. Differentiating the two FOC with respect to r and solving gives

$$\frac{dp/dr}{de/dr} = -\frac{e\ell_e(\ell_e + \ell_p(c - r)) + \ell(\ell_e - \ell_p(c - r))}{\ell_p(2\ell + e\ell_e + e\ell_p(c - r))}$$

[3] This relies on $\epsilon_s^*(\emptyset)$, which is calculated in Section 1.5.

Differentiating this with respect to c gives:

$$\frac{d}{dc}\left(\frac{dp/dr}{de/dr}\right) = \frac{2\ell^2}{\left(2\ell + e(\ell_e + \ell_p(c-r))\right)^2} > 0 \tag{A.19}$$

That is, providers with higher costs adjust in a relatively more price-intensive manner in response to a change in inspection probability. The cost parameter c is therefore identified based on the size of $\frac{dp/dr}{de/dr}$.

A.5.2 Normalizing the scales of effort and κ

The primitives of the model are:

$$\mathcal{P} = \{\{\kappa_s\}, c(e), \{F_\epsilon, F_\theta\}, \{r, \phi\}, \{\alpha, \sigma, \beta, \tau\}\}$$

Proposition A.5.1. *Let $\lambda \in \mathbb{R}^+$. Define*

$$\tilde{\mathcal{P}} \equiv \{\{\lambda\kappa_s\}, c(\lambda e), \{F_\epsilon, F_\theta\}, \{r, \phi\}, \{\alpha, \sigma, \beta, \tau\}\}$$

Any equilibrium $\{p_s^*, e_s^*, \epsilon_s^*(\mathcal{I})\}$ of $\mathcal{P}$ has an observationally equivalent equilibrium $\{p_s^*, e_s^*/\lambda, \epsilon_s^*(\mathcal{I})\}$ in $\tilde{\mathcal{P}}$.

Proof. The proof is almost immediate, as the equations characterizing the building's and provider's optimality conditions are unchanged following this change. □

This proposition says that the scale of κ and e are not separately identified in the data. This is not too worrisome, however, since the counterfactual outcomes of interest are also invariant to this shift in the scales of κ and e. In the empirical work, I use the free normalization on κ to set $\kappa_s = 1$ for the provider for whom I have price data. This allows me to directly infer effort from observed violation frequencies, which is useful when it comes to identifying the cost function.

A.5.3 Finding $\frac{\partial \epsilon_s^*(\mathcal{I})}{\partial p}$

For convenience, I drop the provider subscripts and denote the partial derivative of ϵ^* by ϵ_p^*.

Differentiating Equation 1.2 with respect to p gives:

$$\sigma \epsilon_p^*(\mathcal{I}) = 1 + \sigma\beta \left[\sum_v P(v|\mathcal{I}) \epsilon_p(\mathcal{I}_v)(1 - F_\epsilon(\epsilon^*(\mathcal{I}_v))) \right] \tag{A.20}$$

Define a function $f_0(\mathcal{I})$ and then consider the updating rule

$$f_{n+1}(\mathcal{I}) = 1 + \beta \left[\sum_v P(v|\mathcal{I}) f_n(\mathcal{I}_v)(1 - F_\epsilon(\epsilon^*(\mathcal{I}_v))) \right] \tag{A.21}$$

This updating rule is immediately seen to satisfy Blackwell's sufficient conditions, and therefore defines a contraction. Therefore $\sigma \epsilon_p^*(\mathcal{I})$ is the unique limit of this updating rule. Since the updating rule only depends on the observed cancellation probabilities, $\sigma \epsilon_p^*$ is therefore identified in the data.

Appendix B: Appendix For "Quality Disclosure Design"

B.1 Normalizations of the Probit Model

In its most general form, the inspector's ordered probit model is characterized by the tuple (β, a, b, θ, x). The distribution of a restaurant's inspection scores is by

$$P(s|\beta, a, b, \theta, x) = \Phi\left(\frac{\theta_s - \beta x}{\sqrt{a + bx}}\right) - \Phi\left(\frac{\theta_{s-1} - \beta x}{\sqrt{a + bx}}\right)$$

Under the assumption that b and β are non-zero, the model admits the following normalizations:

- *(Re-scaling of x: $\beta = 1$)* Writing

$$\frac{\theta_s - \beta x}{\sqrt{a + bx}} = \frac{\theta_s - \beta x}{\sqrt{a + \frac{b}{\beta}\beta x}}$$

  gives $P(s|\beta, a, b, \theta, x) = P(s|1, a, b/\beta, \theta, \beta x)$.

- *(Location of latent variable: $a = 0$)* Writing

$$\frac{\theta_s - x}{\sqrt{a + bx}} = \frac{\theta_s + \frac{a}{b} - \left(x + \frac{a}{b}\right)}{\sqrt{b\left(x + \frac{a}{b}\right)}}$$

  gives $P(s|1, a, b, \theta, x) = P\left(s|1, 0, b, \theta + \frac{a}{b}, x + \frac{a}{b}\right)$.

- *(Overall scale: $b = 1$)* Writing

$$\frac{\theta_s - x}{\sqrt{bx}} = \frac{\theta_s/b - x/b}{\sqrt{x/b}}$$

  gives $P(s|1, 0, b, \theta, x) = P\left(s|1, 0, 1, \theta/b, x/b\right)$.

Therefore it is without loss of generality to set $(\beta, a, b) = (1, 0, 1)$ and estimate an ordered probit with heteroskedastic variance satisfying $\text{var}(\epsilon|x) = x$.

B.2 The Grade Share Function

Let $P(g, t|x)$ denote the probability that an inspection cycle results in a grade of g and a re-inspection time of t. For simplicity, I assume the following re-inspection times:

- A on initial inspection: 12 month re-inspection
- Worst of (Initial, Re-Inspection) = B: 6 month re-inspection
- Worst of (Initial, Re-Inspection) = C: 4 month re-inspection

The probability of a given outcome depends on inspector thresholds and the restaurant's latent state x. For example,

$$\begin{aligned} P(A, 12|x) &= P(x + \epsilon < \theta_{13}) \\ &= \Phi\left(\frac{\theta_{13} - x}{\sqrt{x}}\right) \end{aligned}$$

To receive an $(A, 4)$, a restaurant must receive a C on initial inspection, and an A on re-inspection:

$$P(A, 4|x) = \left[1 - \Phi\left(\frac{\theta_{27} - x}{\sqrt{x}}\right)\right]\left[\Phi\left(\frac{\theta_{13} - x}{\sqrt{x}}\right)\right]$$

The remaining states $((A, 6), (B, 6), (C, 4))$ are calculated similarly.

The fraction of time a restaurant has an A grade is given by the expected number of months they'll have an A rating following an inspection cycle divided by the expected number of months following a cycle:

$$s_A(x) = \frac{\sum_t P(A, t|x)t}{\sum_{g,t} P(g, t|x)t}$$

The remaining shares $s_B(x)$ and $s_C(x)$ are calculated analogously.

B.3 More General Heteroskedasticity

In this Section I give support for the assumption of additive heteroskedasticity made in Section 2.5 (see Equation 2.1). In particular, I follow the approach of S. Chen and Khan 2003, which models the ordered probit latent variable in a semiparametric fashion. In this setting, the signal received by the inspector is

$$\hat{x} = x + \sigma(x)\epsilon$$

where $\epsilon \sim \mathcal{N}(0,1)$, and $\sigma(x)$ is an unknown function allowing for arbitrary heteroskedasticity. This model therefore retains the normality assumption, but does not restrict variance to be a linear function of the latent variable x.[1] One difference in the model of Section 2.5 is that x is unobserved, but I show below that this is not a problem.

In Chen and Khan's setup, the key normalizations are setting the smallest and largest thresholds, θ_0 and θ_{S-1}, to 0 and 1, respectively. Writing $P_s(x)$ as the probability of score s given health state x, these normalizations give:

$$P_0(x) = \Phi\left(\frac{-x}{\sigma(x)}\right)$$
$$P_S(x) = 1 - \Phi\left(\frac{1-x}{\sigma(x)}\right)$$

Some manipulation gives

$$\sigma(x) = \frac{1}{\Phi^{-1}(1 - P_S(x)) - \Phi^{-1}(P_0(x))}$$

Thus if $P_0(x)$ and $P_S(x)$ were known, $\sigma(x)$ could be estimated from the data. The value of x and

[1] Other work, such as Lewbel 2000 use an instrumental variables approach to generate moment conditions and relax the normality assumption. However, given the single latent regressor, the conditional independence assumption necessary to adopt a Lewbel-style approach reduces to a standard probit.

the thresholds can then be recovered from the relationships

$$x = -\sigma(x)\Phi^{-1}(P_0(x)) \tag{B.1}$$

$$\theta_s = \sigma(x)\Phi^{-1}\left(\sum_{i \leq s} P_i(x)\right) + x \tag{B.2}$$

Chen and Khan propose using kernel methods to estimate the conditional probabilities $\hat{P}_s(x)$, and then using the plug-in estimator

$$\hat{\sigma}(x) = \frac{1}{\Phi^{-1}(1 - \hat{P}_S(x)) - \Phi^{-1}(\hat{P}_0(x))}$$

Given a long enough panel, you could estimate $P_s(x)$ at a restaurant-level. To increase sample size, I group restaurants according to their mean score, truncated to the nearest half-integer.[2] I then construct kernel density estimates of $\hat{P}_s(x)$, and form plug-in versions of Equations B.1 and B.2.[3]

Figure B.1 plots the resulting estimates. The conditional variance appears to be well-described by an affine function of the latent health state: the coefficient of correlation between the two measures is 0.981. Moreover, the estimated thresholds exhibit a similar kink at the A-B boundary as that seen in Figure 2.7. These findings support the notion that the model of affine heteroskedasticity in Section 2.5 is a good approximation to the true data-generating process.

B.4 Inspector Threshold Heterogeneity

As noted in Section 2.4.4, there is evidence that inspectors vary in their degree of bias towards A's. In this section I show that, in the presence of heterogeneous inspector thresholds, the estimator $\hat{\theta}_s$ from Section 2.6 closely approximates the average of the threshold distribution. The key element of this argument is that inspectors are randomly assigned to restaurants, which implies that (to first order), the observed score distribution at a restaurant is equal to the score distribution of the average inspector.

[2] So long as θ_s is strictly increasing, distinct values of x will generate distinct average scores.

[3] I use a normal kernel with a bandwidth given by Scott's Rule: $h = n^{-\frac{1}{5}}$.

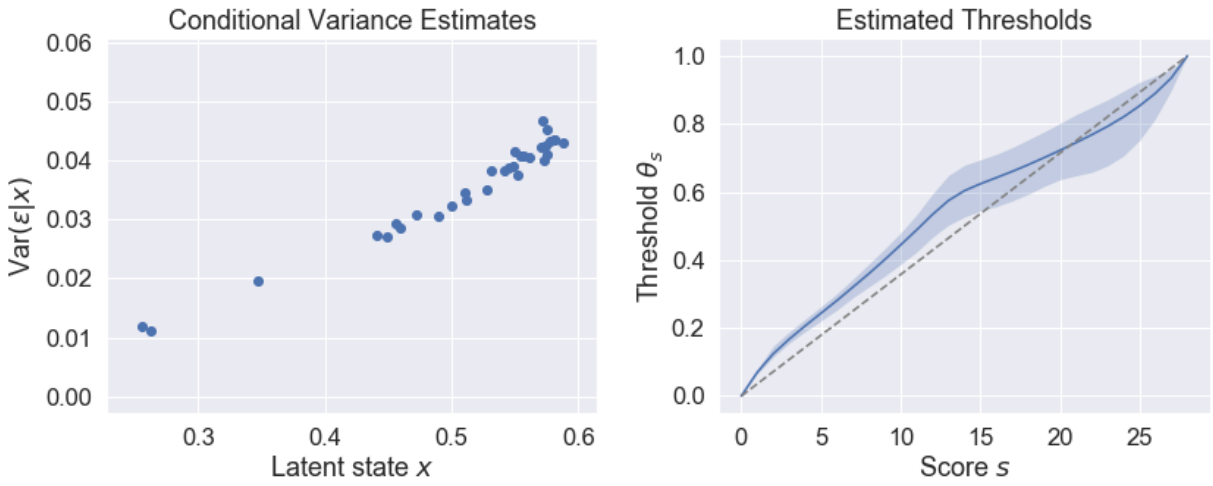


Figure B.1: Heteroskedasticity Estimates using S. Chen and Khan 2003.

Description: Thresholds θ_s are estimated for each restaurant group using Equation B.2. The shaded area in the right-hand chart represents the 5-95% interval of estimated values, and the blue line the average estimated value.

Assume inspector thresholds are given by $\theta^i \sim (\bar{\theta}, \Sigma)$, and assume for simplicity that the restaurant state x is known. The estimated thresholds θ_s in Section 2.6 will, asymptotically, match the following moments for a restaurant of type x:

$$P(s \leq s^*|x) = \Phi\left(\frac{\theta_{s^*} - x}{\sqrt{x}}\right)$$

This implies

$$\hat{\theta}_{s^*} \to_p x + \sqrt{x}\Phi^{-1}\big(P(s \leq s^*|x)\big)$$

With heterogeneous inspector thresholds, random assignment of inspectors implies:

$$\begin{aligned} P(s \leq s^*|x) &= \int \Phi\left(\frac{\theta^i_{s^*} - x}{\sqrt{x}}\right) d\mu(i) \\ &\approx \Phi\left(\frac{\bar{\theta}_{s^*} - x}{\sqrt{x}}\right) + \frac{1}{2}\Phi''\left(\frac{\bar{\theta}_{s^*} - x}{\sqrt{x}}\right)\sigma_s^2 \end{aligned}$$

Thus $\hat{\theta}$ converges in probability to $\bar{\theta}$ plus an error term controlled by the diagonal elements of Σ.

While $\hat{\theta}_s$ is not a consistent estimator of $\bar{\theta}_{s^*}$ under heterogeneity, empirically it is close. To construct Figure B.2, I began with the estimated thresholds θ_s from Section 2.6. From this I extracted two parameters: θ_{30}, the value of θ at the highest score, and $\alpha = \frac{\theta_{13}}{\theta_{30}}$, the fraction of the latent scale in which a restaurant receives an A. I then constructed inspector thresholds using a piecewise linear function:

$$\theta_s^i = \begin{cases} \frac{s\alpha^i\theta_{30}^i}{13} & s \leq 13 \\ \alpha^i\theta_{30}^i + \frac{s-13}{30-13}(\theta_{30}^i - \alpha^i\theta_{30}^i) & s > 13 \end{cases}$$

where

$$\frac{\theta_{30}^i}{\theta_{30}} \sim \mathcal{U}([1-\delta, 1+\delta]), \quad \frac{\alpha^i}{\alpha} \sim \mathcal{U}([1-\delta, 1+\delta])$$

This is a simple way of introducing two types of heterogeneity: (i) all else equal, increasing θ_{30}^i raises each threshold θ_s, which corresponds to a more lenient inspector (lower scores for any value of x); (ii) increasing α^i increases the change in slope before and after the A-B boundary, which corresponds to an inspector with greater clean-kitchen bias.

Figure B.2 shows score distributions (left side) and the probability limit of $\hat{\theta}_s$ for a restaurant with latent state $x = 5$. I've used $\delta = \sqrt{3}/8$, which corresponds to a standard deviation of θ_{30} and α^i equal to one-eighth of their mean.

Most notably, $\hat{\theta}$ is still a good approximation of the average $\bar{\theta}$, particularly for low values of s. The mean absolute percent error is 3.6 percent, with a maximum of 7.3 percent. Interestingly, $\hat{\theta}$ appears to underestimate the change in slope of $\bar{\theta}$, suggesting that the results of Section 2.6 may underestimate the extent of A grade bias if inspector heterogeneity is substantial.

B.5 OATH Tribunals

As noted in Section 2.2, restaurants that do not receive an A upon re-inspection can challenge inspection findings through the city's Office of Administrative Trials and Hearings ("OATH"). If

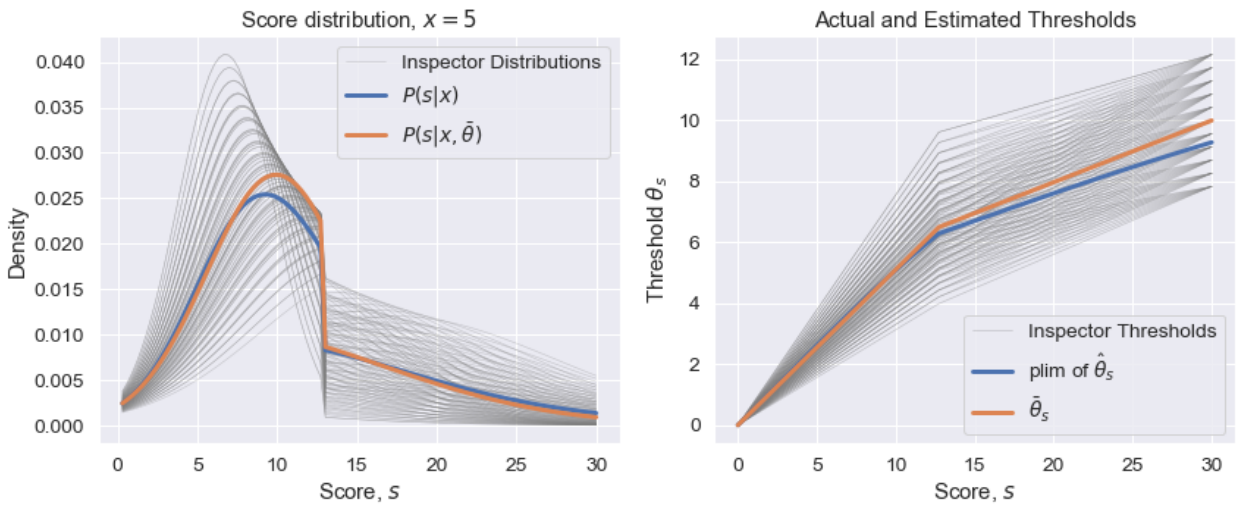


Figure B.2: Score Distributions and Estimated Thresholds under Inspector Heterogeneity.

Description: Grey lines represent individual inspector score distributions for a restaurant with latent state $x = 5$ and latent thresholds, respectively. Blue lines represent the probability limit of the observed score distribution and $\hat{\theta}$, respectively. Orange lines represent the score distribution of an inspector with $\theta^i = \bar{\theta}$ and the average threshold, $\bar{\theta}$, respectively.

the OATH board dismisses enough violations to result in a new grade, the restaurant receives a new grade card corresponding to the improved grade. It is not uncommon for an OATH hearing to result in a new grade: as shown in Figure B.3, between 2011 and 2016, 20-40% of B grades were overturned to A's through an OATH hearing.

The scores in the primary sample are post-OATH scores. Given that OATH hearings will only ever improve a restaurant's score, this raises concerns about whether the results presented in Sections 2.6 and 2.7 are influenced by OATH hearings. For example, is it the OATH process, and not inspector preferences, that explains the score-bunching noted in Figure 2.3?

To address this question, I take advantage of the 2011-16 dataset that includes both pre- and post-OATH scores for each inspection. Figure B.4 shows the results of the primary estimation procedure using each set of scores. Notably, the estimated thresholds θ_s are not particularly sensitive to the choice of score. If anything, the kink in the threshold around the A-B boundary is more pronounced when estimated using pre-OATH scores, suggesting that the estimates using post-OATH

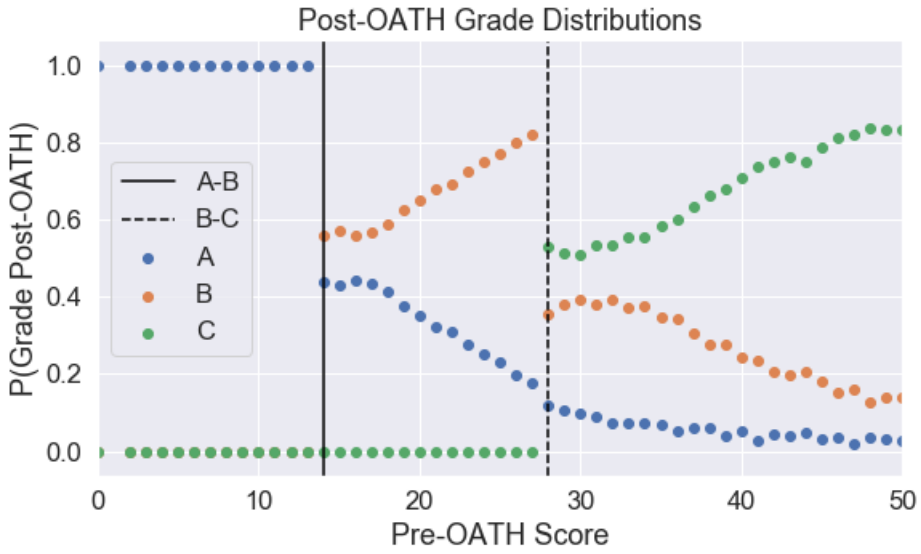


Figure B.3: Post-OATH Grade Distributions, 2011-16 Inspection Data.

Description: Empirical grade distributions based on 2011-16 inspection data.

data may understate the extent to which inspectors contribute to score bunching.

The biggest difference between the two estimates is the restaurant states. Since post-OATH scores are no worse than pre-OATH scores, the estimation routine assigns restaurants lower levels of x (cleaner states) when using post-OATH data. The median estimated value of x when using pre- and post-OATH data is 5.11 and 4.57, respectively.

OATH hearings also influence a restaurant's first-order condition, since they affect the distribution of grades a restaurant receives conditional on x. Let $P_{oath}(g|s)$ be the probability that a restaurant receives a post-OATH grade of g given a pre-OATH score of s. Then, for example, the probability that a restaurant is in an A-6 state is:

$$P(A-6|x) = \underbrace{P_I(B|x)}_{\text{B on initial}} \left[\underbrace{P_I(A|x)}_{\text{A on re-insp}} + \underbrace{\sum_{s=14}^{27} P_I(s|x) P_{oath}(A|s)}_{\text{B on re-insp + OATH overturn}} \right]$$

These probabilities can be translated into a grade share function using the approach in Section B.2.

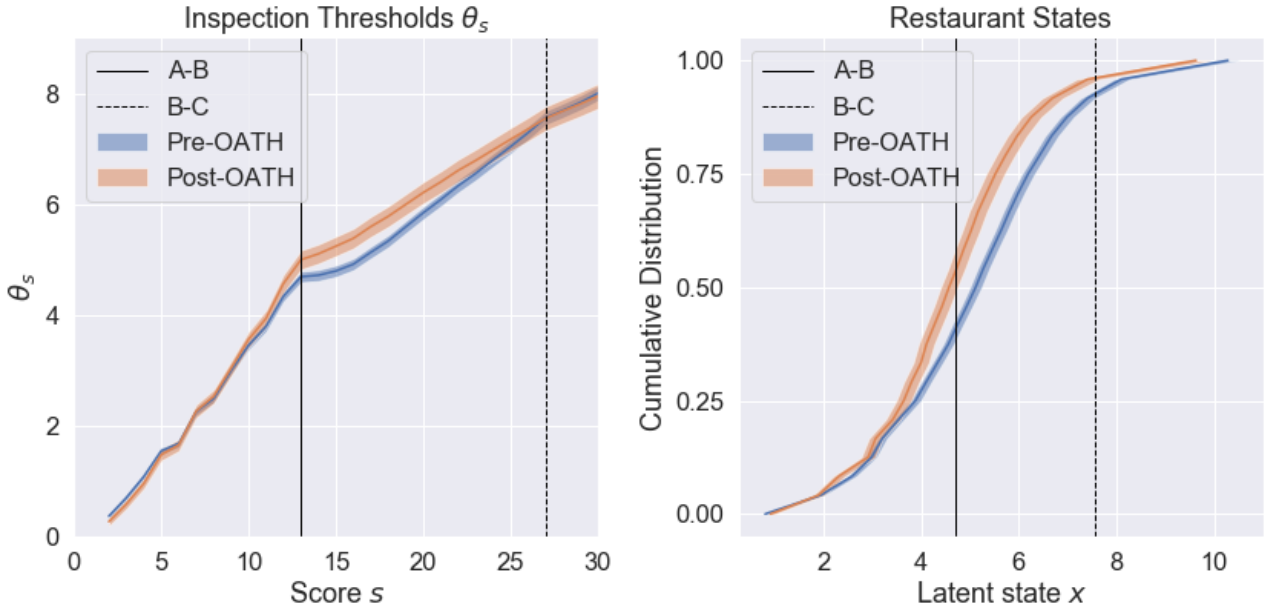


Figure B.4: Estimation Results with pre- and post-OATH Data, 2011-16.

Description: Estimated values of thresholds θ_s and restaurant states x_r. Confidence intervals based on plug-in estimate of MLE asymptotic variance.

To understand how much these differences influence counterfactual outcomes, I re-estimate the counterfactual from Section 2.7.3 (removing re-inspections) as follows. Since the 2011-16 sample does not contain matched food-poisoning data, I do this by adjusting the estimates from the 2017-19 dataset in two ways. First, I approximate the x-values I would had estimated if the 2017-19 data included pre-OATH scores. Define x_r as the estimated state for restaurant r in the primary dataset; $q^{pre}(\alpha)$ and $q^{post}(\alpha)$ as the α-th quantiles of the 2011-16 restaurant states when estimated on pre- and post-OATH data, respectively; and $F(x)$ the CDF of estimated restaurant states in the primary sample. The adjusted value of x is:

$$\tilde{x}_r \equiv x_r \frac{q^{pre}(F(x))}{q^{post}(F(x))}$$

Second, I adjust the restaurant's first-order condition to explicitly account for the OATH process.

In this scenario, I estimate a 7.0-9.5% drop in food poisoning cases as a result of removing

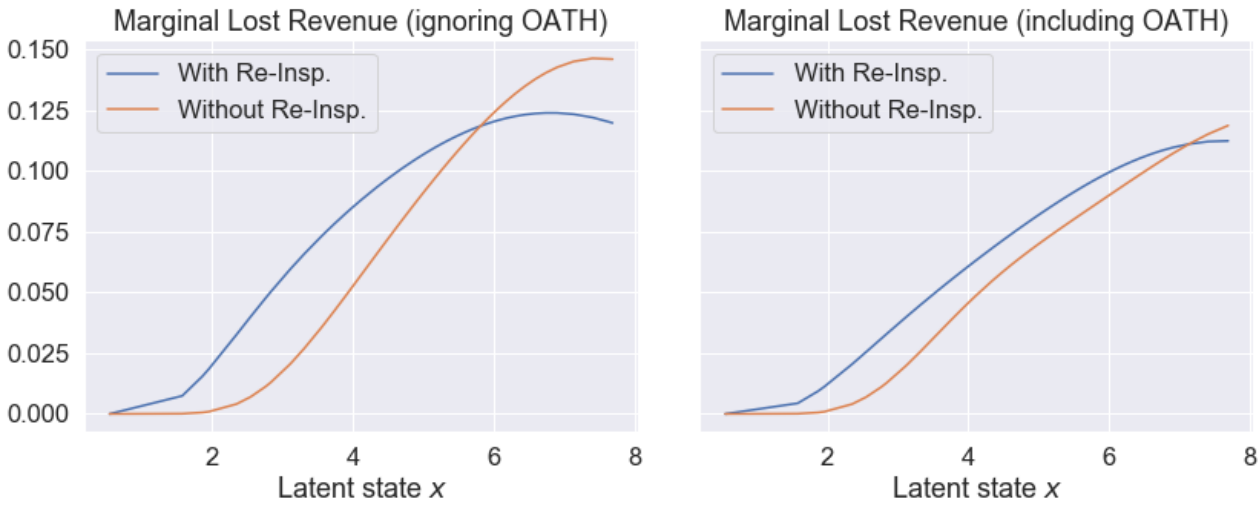


Figure B.5: First-Order Conditions with and without OATH Hearings.

Description: The left-hand figure plots the marginal lost revenue for the baseline model with (blue) and without (orange) re-inspections. In particular, it shows $-(s'_A(x) + s'_B(x)\tilde{\rho})$ for $\tilde{\rho} = 0.5$. The right-hand figure shows the same quantities when the share function $s_g(x)$ has been modified to account for the OATH process.

re-inspections, as compared to a 9.8 to 14.4% drop in the baseline results. The reason the estimate decreases is that OATH hearings act as a buffer: as shown in Figure B.5, the marginal effect of removing re-inspections on restaurant revenue is less pronounced when they still have OATH hearings as an avenue for improving their grades. This causes predicted improvements in x to decreases, leading to a smaller decline in food-poisoning cases.

In summary, ignoring the OATH process in the main results appears to have little impact on estimated inspector thresholds, but cause the counterfactuals to overstate the magnitude of food-poisoning declines. However, the qualitative conclusions appear to remain unchanged when accounting for OATH hearings.

B.6 More General Cost Functions

In the baseline model of Section 2.5, investment costs are a decreasing, linear function of the health state x:

$$c_r(x) = C_r \cdot (\bar{X} - x)$$

However, there is no a-priori reason to expect investment costs to be linear. For instance investment costs could be convex in x: for very clean restaurants the marginal cost of improving is quite high, whereas dirtier restaurants may have access to relatively simple, low-cost improvements.

Without imposing any functional form restrictions, and taking $\tilde{\rho}_r = 0$ to simplify the analysis, the restaurant's first-order condition is

$$-\frac{c_r'(x)}{\Delta_{rA}} = -s_A'(x),$$

which says that at an optimum, the marginal cost of lowering x needs to balance the associated revenue gain. The first-order condition identifies the restaurant's marginal cost, relative to its loss from falling from an A to C grade.

The shape of the restaurant's cost curve determines its response to a counterfactual change in policy. Convex cost curves will tend to produce smaller changes in x, as shown in Figure B.6. In this example, a restaurant chooses $x_0 = 4$ under the baseline policy, and some counterfactual policy (such as lowering the A-B boundary) causes the marginal lost revenue curve to shift to the left. If costs are linear, the restaurant's new optimal point is x_c^{lin}, whereas if costs are convex, the new optimal point is x_c^{con}. The improvement in the health state is smaller when costs are convex, since the marginal cost savings from reducing x are decreasing. Thus if restaurants face strongly convex investment costs, the results in the main chapter may overstate the potential gains from counterfactual policies.

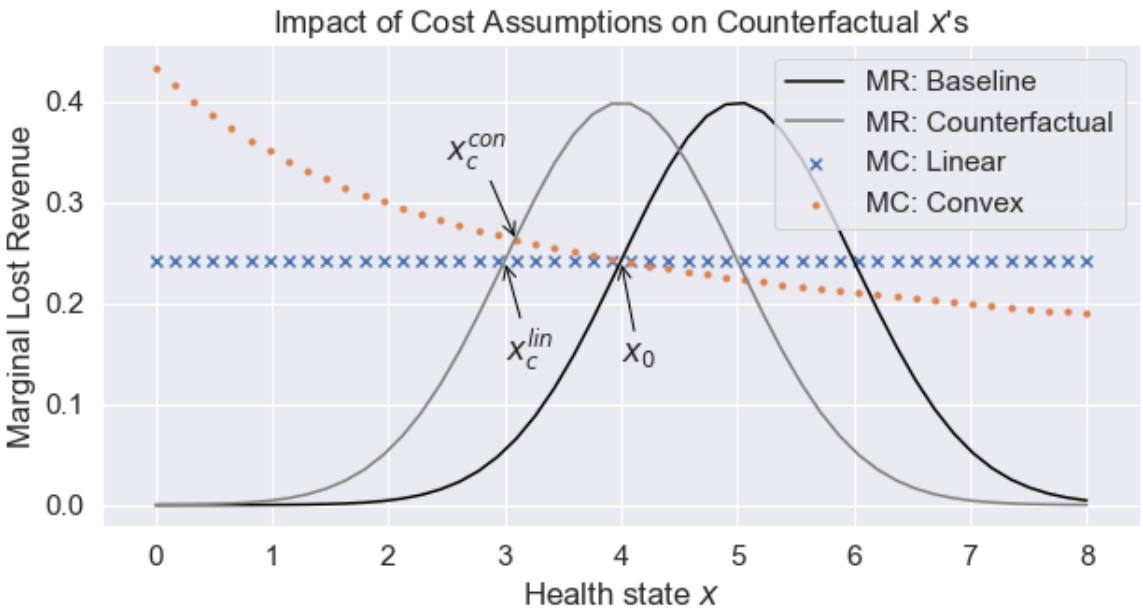


Figure B.6: First-Order Conditions for Various Cost Functions.

Description: Illustration of the restaurant's first-order condition for linear (blue) and convex (orange) cost functions. The improvement in x is smaller when costs are convex.

Appendix C: Appendix for "Real-Time Inference"

C.1 Identifying Statistical Discrimination

The model of Section 3.5 assumes umpires are statistical decision-makers and shows that, under this assumption, there is little evidence of taste-based discrimination (that is, variation in the relative cost of false negatives by count). Of course, umpire behavior could be consistent with pure taste-based discrimination, with their tendency to expand the strike zone in high-ball counts driven by costly false negatives in such counts. In this appendix I discuss how the shape of the umpire's strike zone can, in theory, identify the extent of taste-based versus statistical discrimination in the model. While the empirical findings are inconclusive, I believe this is a unique method for separating types of discrimination.

To circumscribe the problem, I assume in count C the umpire's prior over pitch locations is

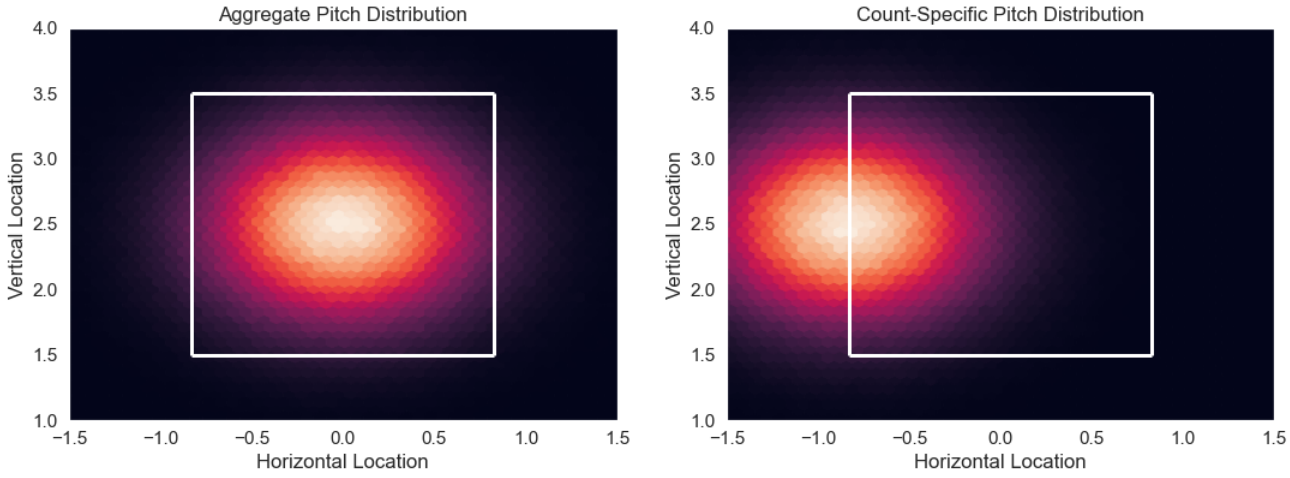
$$P^u(x, z|C) \equiv \alpha P(x, z) + (1 - \alpha)P(x, z|C) \tag{C.1}$$

where $P(x, z)$ is the aggregate probability of a pitch at location (x, z), and $P(x, z|C)$ the count-specific probability. The parameter α controls the extent of statistical discrimination. When $\alpha = 0$, umpires are perfect statistical discriminators, whereas when $\alpha = 1$ umpires do not update their priors at all in response to count information.[1] While this assumption substantially restricts the types of priors umpires are allowed to hold, umpires that hold vastly different beliefs — and therefore call vastly different strike zones — would presumably have been filtered out of the system before reaching the Major Leagues.

How is α identified from umpire ball/strike calls? Consider the setup shown in Figure C.1:

[1] In the empirical work I restrict $\alpha \in [0, 1]$, but there is no a priori reason to do so: $\alpha < 0$ would simply reflect an umpire that overreacts to count information, and $\alpha > 1$ an umpire whose prior updates in the wrong direction.

Figure C.1: Hypothetical Pitch Distributions



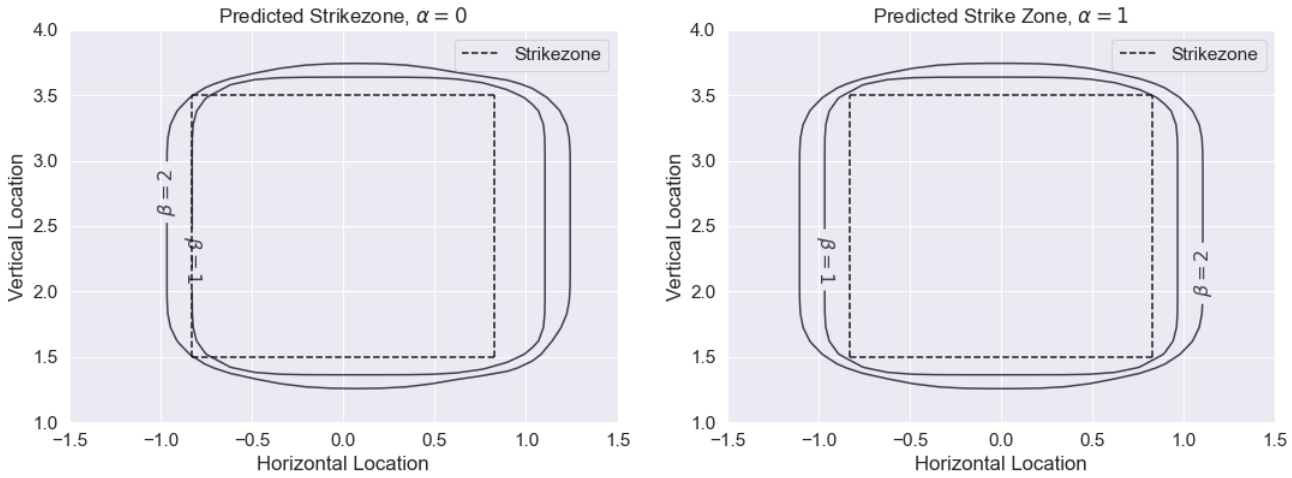
Description: Two hypothetical pitch distributions. Lighter areas represent regions of higher density.

in aggregate, pitches are normally distributed around the center of the strike zone, but in count C pitches are distributed around the left boundary of the zone. In Figure C.2 I plot the "median" strike zone in count C for umpires with different values of α and β — that is, the black lines denote the boundary of the region in which umpires call a strike more than 50% of the time.

When umpires are statistical discriminators ($\alpha = 0$), their strike zone shifts about two inches to the right in count C. This is because, at the righthand boundary, the pitch density is decreasing as you cross the boundary, so the umpire's prior pulls pitches on the righthand boundary back towards the strike zone. However, since the pitch distribution is symmetric around the lefthand boundary, the prior exerts little pull in this setting, and the umpire's strike zone lines up with the actual zone.

Importantly, changing β cannot generate a rightward shift of the strike zone in count C. As β increases, mistakenly calling a ball becomes more costly no matter the location of the pitch, and so the umpire's strike zone simply scales radially. Therefore α could in principle be identified by changes in the shape of the called strike zone across counts: a rightward shift would be evidence of statistical discrimination, whereas no shift (or a leftward shift) would be evidence of no (or incorrect) statistical discrimination.

Figure C.2: Predicted Strike Zone in Count C



Description: This figure plots an umpire's median strike zone (the region in which a pitch is called a strike more than half of the time) for the count-specific pitch distribution shown on the right-hand side of Figure C.1. The left side shows the strike zone for a perfect statistical discriminator ($\alpha = 0$), and the right side shows the strike zone for an umpire whose prior never updates ($\alpha = 1$).

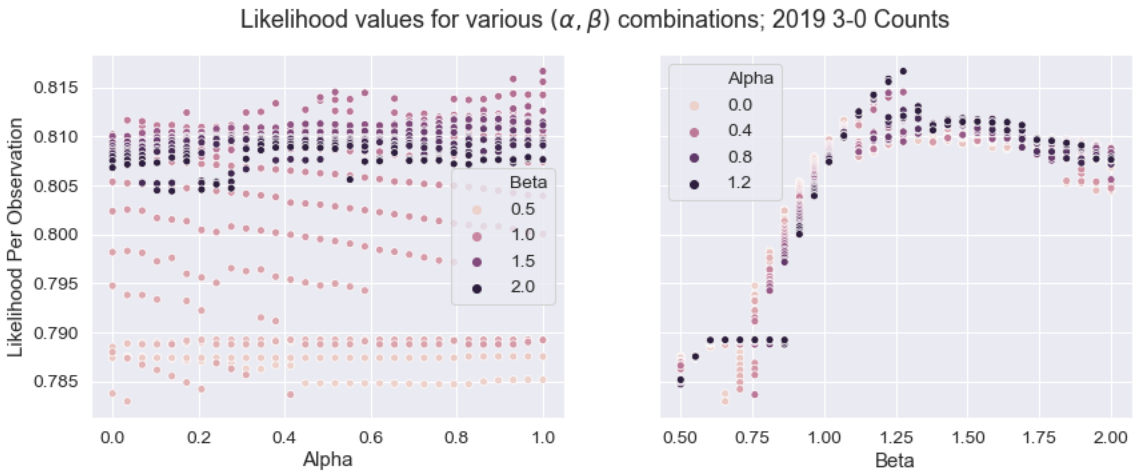
The crucial feature for identifying α in this model is that the cost parameter is lower-dimensional than the umpire's signal. If the cost of a false-negative was higher for pitches on the right-hand side of the plate, then a rightward shift could be consistent with no statistical discrimination.

Unfortunately, estimating this extended model on 2019 data does not yield a clear prediction on the value of α. Let $\mathcal{L}(\sigma_x, \sigma_z, \beta, \alpha)$ represent the log-likelihood generated by Equation 3.2 with the prior given by Equation C.1. Define the marginalized log-likelihood:

$$\mathcal{L}^*(\beta, \alpha) = \max_{\sigma_x, \sigma_z} \mathcal{L}(\sigma_x, \sigma_z, \beta, \alpha)$$

In Figure C.3 I plot $\mathcal{L}^*(\beta, \alpha)$ on a 30×30 grid of values for β, and α. While the likelihood is noticeably quadratic in β around $\beta \approx 1.25$, the likelihood is largely flat as a function of α. This is perhaps not too surprising: α will be better identified when the pitch distribution is highly asymmetric about the origin, but none of the distributions in Figure 3.4 exhibit rotational asymmetry as severe as Figure C.1.

Figure C.3: Maximum Likelihood Estimates for (β, α)



Description: This figure shows two views of $\mathcal{L}^*(\beta, \alpha)$ evaluated using 2019 data in 3-0 counts. The left side plots α along the x-axis, and each vertical dot represents a different value of β, and the right side vice versa. The log-likelihood has been transformed into units of likelihood per observation: $\exp(\mathcal{L}^*(\beta, \alpha)/N)$.

As discussed in Bohren et al. 2019, distinguishing taste-based from statistical discrimination in a context where decision-makers may have incorrect beliefs is difficult. The technique of using the contours of the decision-maker's acceptance set to identify statistical discrimination is, to the best of my knowledge, new. However, as evidenced above, there are some challenges to using this technique. Contour-based inference is most likely to work when: (i) historical data and selection standards put bounds on the beliefs a decision-maker can hold; (ii) the econometrician has sufficient data to accurately estimate the decision-maker's contours and avoid omitted variable bias; and (iii) conditional on group status, the degree of taste-based animus is independent of observables.